RICE UNIVERSITY

Redundancy-aware learning of protein structure-function relationships

by

Drew Bryant

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctorate of Philosophy

APPROVED, THESIS COMMITTEE:

Lydia E. Kavraki, Chair
Noah Harding Professor of Computer Science and Bioengineering

Luay Nakhleh
Associate Professor of Computer Science

Yousif Shamoo
Associate Professor of Biochemistry and Cell Biology

Houston, Texas

April, 2012

ABSTRACT

Redundancy-aware learning of
protein structure-function relationships

by

Drew Bryant

The protein kinases are a large family of enzymes that play a fundamental role in propagating signals within the cell. Because of the high degree of binding site similarity shared among protein kinases, designing drug compounds with high specificity among the kinases has proven difficult. However, computational approaches to comparing the 3-dimensional geometry and physicochemical properties of key binding site residues, referred to here as *substructures*, have been shown to be informative of inhibitor selectivity. This thesis introduces two fundamental approaches for the comparative analysis of substructure similarity and demonstrates the importance of each method on a variety of large protein structure datasets for multiple biological applications.

The Family-wise Alignment of SubStructural Templates Framework (The FASST Framework) provides an unsupervised learning approach for identifying substructure clusterings. The substructure clusterings identified by FASST allow for the automatic evaluation of substructure variability, the identification of distinct structural conformations and the selection of anomalous outlier structures within large structure datasets. These clusterings are shown to be capable of identifying biologically meaningful structure trends among a diverse number of protein families. The FASST Live visualization and analysis platform provides multiple comparative analysis pipelines and allows the user to interactively explore

the substructure clusterings computed by The FASST Framework.

The Combinatorial Clustering Of Residue Position Subsets (CCORPS) method provides a supervised learning approach for identifying structural features that are correlated with a given set of annotation labels. The ability of CCORPS to identify structural features predictive of functional divergence among families of homologous enzymes is demonstrated across 48 distinct protein families. The CCORPS method is further demonstrated to generalize to the very difficult problem of predicting protein kinase inhibitor affinity. CCORPS is demonstrated to make perfect or near-perfect predictions for the binding ability of 12 of the 38 kinase inhibitors studied, while only having overall poor predictive ability for 1 of the 38 compounds. Additionally, CCORPS is shown to identify shared structural features across phylogenetically diverse groups of kinases that are correlated with binding affinity for particular inhibitors; such instances of structural similarity among phylogenetically diverse kinases are also shown to not be rare among kinases. Finally, these function-specific structural features may serve as potential starting points for the development of highly specific kinase inhibitors.

Importantly, both The FASST Framework and CCORPS implement a *redundancy-aware* approach to dealing with structure overrepresentation that allows for the incorporation of all available structure data. As shown in this thesis, surprising structural variability exists even among structure datasets consisting of a single protein sequence. By incorporating the full variety of structural conformations within the analysis, the methods presented here provide a richer view of the variability of large protein structure datasets.

Acknowledgments

I would like to express my deepest appreciation for the enormous amount of both career and life advisement provided by my adviser, Dr. Lydia Kavraki. I would also like to thank Dr. Luay Nakhleh and Dr. Yousif Shamoo for their encouragement and insightful comments while pursuing the work presented in this thesis.

I am greatly indebted to Dr. Brian Chen for the *many* hours that he chose to invest in personally teaching and training me over the several years that we worked together. Finally, I would like to thank Dr. Mark Moll and Dr. Paul Finn for their guidance and assistance in the work developed in this thesis.

This work was supported by National Science Foundation Graduate Research Fellowship grant DGE-0237081 to DHB, NSF ABI grant ABI-0960612, and the John and Ann Doerr Fund for Computational Biomedicine at Rice University (Texas Higher Education Coordinating Board NHARP 01907). Equipment used was supported by National Science Foundation grants CNS-0454333 and CNS-0421109 in partnership with Rice University, AMD and Cray. This work was also supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

Contents

Illustrations

Tables

Chapter 1

Introduction

1.1 Motivation

The disruption of signaling networks within the cell has been implicated in a wide variety of disease states [7, 8]. Many of the signal propagating agents within these networks belong to a single, large family of enzymes—the protein kinases. Because of the highly interconnected nature of the human kinase signaling network, even the dysregulation of a single protein kinase, such as p38α, can lead to multiple pathological conditions. Specifically, p38α dysregulation has been implicated in tumor formation, cell cycle disruption, and inflammation disorders, such as rheumatoid arthritis and Alzheimer's [8].

Because of the number of kinase-associated conditions, the protein kinases have come to constitute 20-30% of the drug development programs at many companies [9]. However, designing highly targeted kinase inhibitors has proven difficult for several reasons. Firstly, the primary binding site targeted by approved kinase inhibitors, the ATP binding pocket, happens to have considerable structural similarity across protein kinase domains [10, 11]. Secondly, because the protein kinases constitute the largest protein family encoded by the

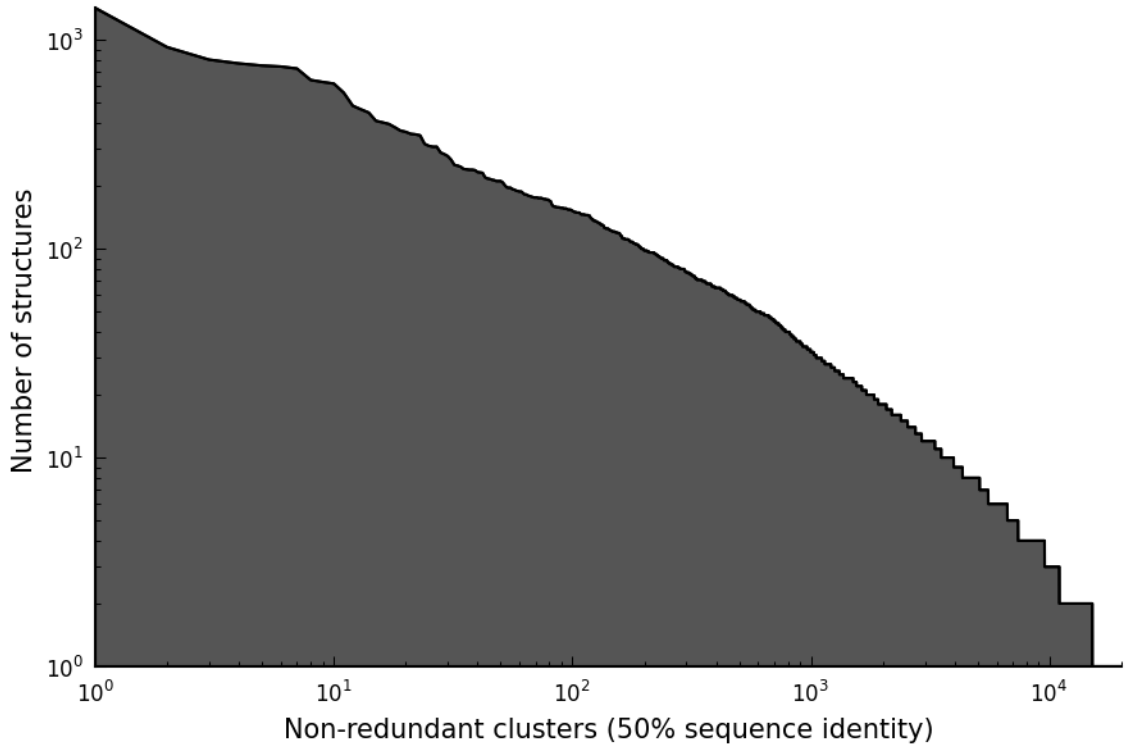


Figure 1.1 : **Structural overrepresentation within the PDB.** Log-log plot of the number of available structures per non-redundant sequence cluster at 50% identity (NR_{50}-clusters) for the entire PDB. As shown above, the distribution of structures among NR_{50}-clusters within the PDB is highly non-uniform. Of the total 19975 NR_{50}-clusters, 198 contain 100 or more structures while 4883 contain only a single structure.

human genome [9, 4], a large number of potential off-targets for any given kinase inhibitor naturally exist. Therefore, the design of highly targeted kinase inhibitors necessitates the comparative structural analysis of binding sites to identify differentiating structural features [11].

Part of the difficulty in identifying differentiating structural features among large structure families, such as the protein kinases, is the difficulty of defining useful similarity measures that account for the structural features of importance. To quantify the *functional* similarity of binding sites, *local structural* similarity has proven to be a highly informative

proxy that is more directly quantifiable [12, 13, 14, 15]. For example, the 3-dimensional position and physicochemical properties of binding site residues, such as the size of the "gatekeeper" residue that mediates the availability of the kinase ATP binding site to inhibitors, have been shown to be highly informative of inhibitor binding specificity [10, 11]. Comparison of the key binding site residue positions (local structural features), rather than the overall protein fold or topology (global structural features), has been demonstrated to differentiate among the binding ability of different inhibitors [10, 11, 16].

A major obstacle for the development of local structure comparative methods is the problem of structural overrepresentation in the Protein Data Bank (PDB) [17]. As shown in Fig. 1.1, the distribution of protein structures in the PDB is highly non-uniform. Currently, 4883 protein sequences have only a single structure with $\geq$50% sequence identity and yet, at the same level of sequence identity, many sequences have over 100 available structures. Targets of pharmaceutical development, such as p38α, exhibit extreme overrepresentation—p38α alone currently has 160 available structures. Because of the degree of overrepresentation, identifying trends among datasets having a highly non-uniform structure distribution across proteins becomes inherently difficult. Without accounting for the overrepresentation, the trends identified will be largely trends among the most prevalent structures, representing only a fraction of the proteins of interest, rather than among the unique proteins themselves.

A typical approach to dealing with overrepresentation is to construct sequence non-redundant subsets of the original dataset and perform the analysis on now uniformly distributed structural data. However, as will be demonstrated throughout this thesis, the selection of a single structural representative for a given protein sequence is highly non-trivial due to the fact that many proteins exhibit multiple structurally distinct conformations that cannot be simultaneously captured by a single representative structure.

Furthermore, as the number of available 3-dimensional protein structures in the PDB continues to grow, as well as the frequency of overrepresentation, identifying meaningful trends among the structures becomes increasingly difficult. Identifying structural trends, specifically *local structural* trends, such as instances of binding site residue similarity, has been shown to be highly predictive of the biological function of those binding sites [18, 13, 15, 16]. The term *substructure* is used throughout this thesis to refer to such sets of 3-dimensionally grouped residues that may or may not be sequence contiguous.

1.2 Contributions

This thesis introduces two fundamental approaches for the comparative analysis of substructure similarity and demonstrates the importance of each method on a variety of large protein structure datasets for several biological applications.

- The Family-wise Alignment of SubStructural Templates Framework (The FASST Framework) provides an unsupervised learning approach for identifying substructure clusterings. The substructure clusterings identified by FASST allow for the automatic evaluation of substructure variability, the identification of distinct structural conformations and the selection of anomalous outlier structures within large structure datasets. The substructure clusterings identified by FASST both here and previously [1] are demonstrated to be capable of identifying biologically meaningful structure trends among a diverse number of protein families. Importantly, the implementation of The FASST Framework provides a modular set of components that allow for the quantification of local structure similarity between substructures to be easily customized for different biological applications.

- The FASST Live visualization and analysis platform provides multiple comparative

analysis pipelines and allows the user to interactively explore the substructure clusterings computed by The FASST Framework without the need to manually develop a custom pipeline from The FASST Framework components.

- The Combinatorial Clustering Of Residue Position Subsets (CCORPS) method provides a supervised learning approach for identifying structural features that are correlated with a given set of annotation labels. No assumptions regarding the nature of the annotation labels nor of the alignment type are made by CCORPS. The generality of CCORPS allows for many different types of biological applications to be addressed without requiring modification to the underlying algorithm. The ability of CCORPS to identify structural features predictive of functional divergence among families of homologous enzymes is demonstrated across 48 distinct protein families. The strong performance of CCORPS across the families demonstrates the generality of the approach for protein families that vary greatly in the number of available structures, the number of functional classes and the number of binding site residue positions.

- The CCORPS method is further demonstrated to generalize to the very difficult problem of predicting protein kinase inhibitor affinity. The protein kinase affinity experiments presented in this thesis demonstrate the largest, both in number of structures and number of inhibitors considered, structural analysis of the human kinome to date. The predictive ability of CCORPS is compared to a recent state-of-the-art structure-based approach and CCORPS is demonstrated to meet or exceed the capabilities of previous methods. Specifically, CCORPS is demonstrated to make perfect or near-perfect predictions for the binding ability of 12 of the 38 kinase inhibitors tested in the inhibitor affinity dataset presented, while only having overall poor predictive ability for 1 of the 38 compounds. Additionally, CCORPS is shown to identify shared

structural features across phylogenetically diverse groups of kinases that are correlated with binding affinity for particular inhibitors; such instances of structural similarity among phylogenetically diverse kinases are also shown to not be rare among kinases. These function-specific structural features may serve as potential starting points for the development of highly specific kinase inhibitors.

Finally, both The FASST Framework and CCORPS implement a *redundancy-aware* approach to dealing with structure overrepresentation that allows for the incorporation of all available structure data without requiring sequence non-redundant pre-filtering. As illustrated by The FASST Framework, surprising structural variability exists even among structure datasets consisting of only a single protein sequence and incorporating the full variety of structural conformations within the analysis presented here provides a richer view of the variability of protein structure datasets.

1.3 Thesis overview

First, important terminology and concepts that are referred to throughout the thesis, such as the distinctions between unsupervised and supervised learning and the basis for structure-function relationships are presented in Chapter 2. Chapter 3 introduces both The FASST Framework and FASST Live and demonstrates the generalizability of the substructure clustering approach to many different problem domains. Chapter 4 introduces the CCORPS method itself, and Chapters 5 and 6 demonstrate the predictive ability of CCORPS for a variety of large protein structure datasets.

Chapter 2

Background

In order to provide a minimal foundation for the remainder of the thesis, an overview of the major distinctions between unsupervised and supervised learning is discussed here as well as a brief introduction to structure-function relationships as they relate to protein enzymatic sites [19].

2.1 The unsupervised learning problem

The problem of unsupervised learning pertains to identifying the underlying structure or patterns that exist within an *unlabeled* dataset. Notable applications include ([20] pp.517–518):

- Natural grouping identification (clustering)
- Novelty detection
- Outlier and anomaly detection
- Trend identification

- Compression (data reduction)
- Data uniformity assessment

In particular, the work presented in this thesis will make use data clustering throughout to better understand the natural similarity groups that occur among 3-dimensional protein structures.

Ideally, items falling within the same cluster should have smaller distances to one another than to items in other clusters. Therefore, a precise definition of item-item distance is necessary in order to quantify the quality of a cluster. In fact, the notion of a cluster is so intertwined with a specific definition of the item-item distance that completely different clusterings can be obtained by simply altering the distance function.

Take, for example, a bag containing spheres having `small`, `medium` and `large` sizes that are colored either `red`, `green` or `blue`. Assume that at least one of each color exists for each size. Given distance measures $d_{color}(x,y) = \{0 \text{ if color}(x) = \text{color}(y), 1 \text{ otherwise}\}$ and $d_{size}(x,y) = \{0 \text{ if size}(x) = \text{size}(y), 1 \text{ otherwise}\}$. Applying an agglomerative hierarchical clustering method ([20] pp.552–553) and stopping at the level of 3 clusters while using d_{color} will result in color-homogeneous clusters, but within each cluster will be a mixture of all sizes. Repeating the clustering using d_{size} instead will result in size-homogeneous clusters, each containing a mixture of all colors. Finally, repeating the clustering again using $d_{both} = d_{color} + d_{size}$ and stopping at 9 clusters would result in clusters homogeneous in both size and color.

2.2 The supervised learning problem

The problem of supervised learning is distinguished from unsupervised learning by the addition of a set of *labels* to the data points. The addition of labels allows supervised

approaches to identify trends or groupings within the data that correlate with the given set of labels. Most applications of supervised learning involve using a labeled training set to learn a set of rules that can then be applied to predict the labels for novel, unlabeled data ([20] pp.16–17).

For example, the problem of spam detection is easily cast as a supervised learning problem. Given a set of emails (data points) that have been labeled as either `spam` or `not-spam` (labels), train a classifier to learn the features of the emails (e.g., email length, prevalence of CAPITILIZATION, number of misspelled words). Then, given a new (unlabeled) email message, the classifier applies the learned decision rules to predict the label for the new email—`spam` or `not-spam`.

2.3 Structure-function relationships

The function of a protein is intimately linked to its 3-dimensional structure [14, 21]. Specifically, the 3-dimensionally grouped residues (amino acids) that make up the binding sites of enzymatic proteins have been shown in several instances to dictate function alone [22]. That is, given two proteins without discernible sequence or fold similarity, *substructure* approaches have been shown capable of identifying local instances of structural similarity indicative of functional similarity [23]. The term *substructure* is used throughout this thesis to refer to such sets of 3-dimensionally grouped residues that may or may not be sequence contiguous.

For example, the `Ser-His-Asp` catalytic triad of the serine proteases is a well-understood example of function-determining substructure that appears to have been convergently evolved in both bacterial and mammalian species. The 3-dimensional geometry of the triad is shared among both lineages despite the absence of both sequence and topological similarity.

Because of the strong relationship between the 3-dimensional structure and function of proteins, repositories such as the Structure-Function Linkage Datbase (SFLD) [24] have been developed to provide curated annotations. For example, the Enolase Superfamily (ES) is currently known to include 20 different enzymatic protein families that share a common partial reaction. The shared partial reaction is performed using a common set of active site residues in each family and has been heavily conserved throughout the course of evolution in spite of specialization of the remainder of the binding site. For example, one member of the ES is depicted in Fig. 2.1 and the 5-residue catalytic substructure that has been conserved throughout the ES is shown.

The Enzyme Commission (EC) class annotation [25] for a given enzyme provides a hierarchical classification of the type of reaction catalyzed by the enzyme as well as the type of substrate molecule on which the enzyme operates. In the 4-tiered EC classification `A.B.C.D`, where each level is delimited by a period (`.`), the values for `A` through `D` increase in annotation specificity. For example, the structure of mandelate racemase (Fig. 2.1) has been annotated to have the EC class `5.1.2.2`. The definitions for each of the 4 tiers is:

EC Number	Definition
`5.-.-.-`	`isomerase`
`5.1.-.-`	`racemase or epimerase`
`5.1.2.-`	`acting on hydroxy acids and derivatives`
`5.1.2.2`	`mandelate racemase`

where `isomerase` is the least specific functional annotation and the `mandelate racemase` annotation is the most specific, providing both the reaction catalyzed and the preferred substrate molecule of the enzyme.

The functional diversification of the ES can be examined by comparing the EC class annotations of the proteins known to be ES members. A subset of the reactions catalyzed

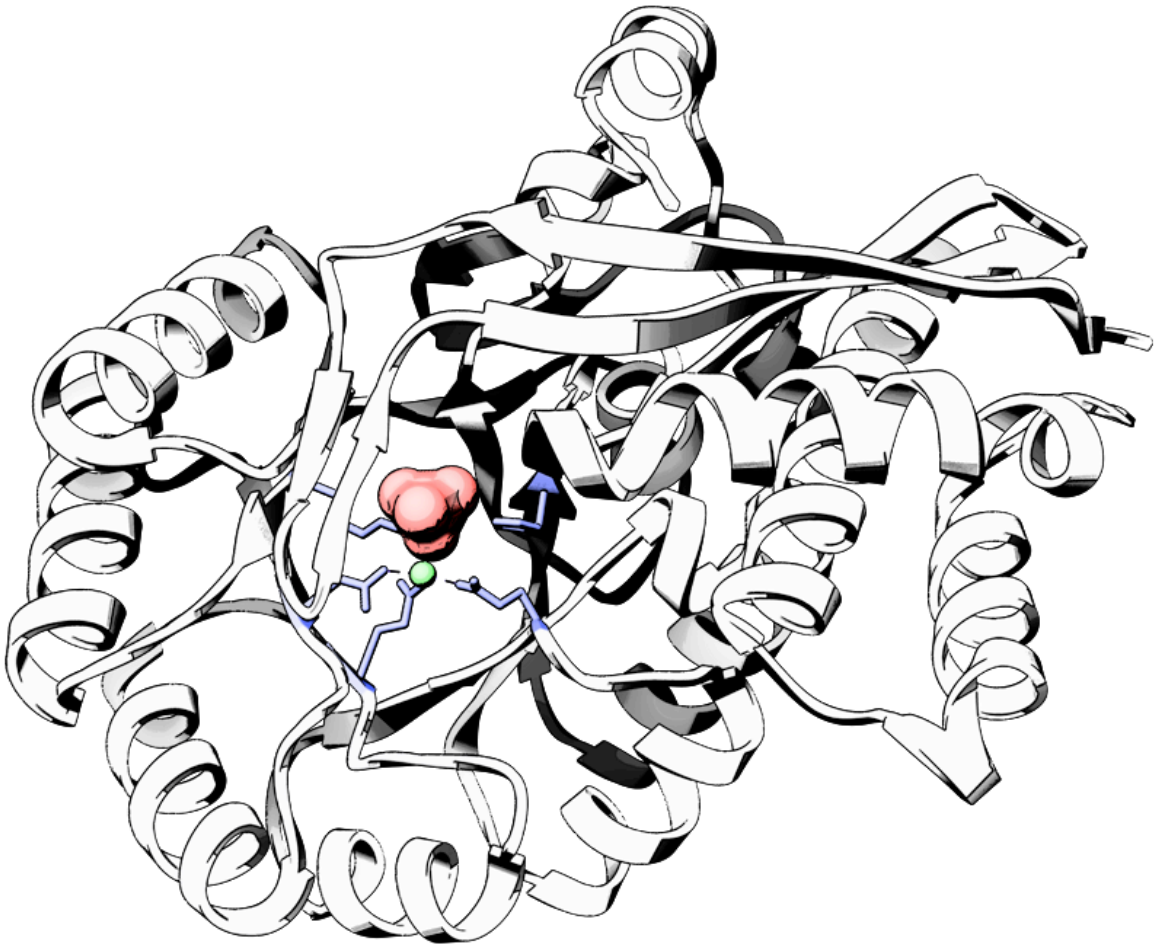


Figure 2.1 : **Mandelate racemase active site.** The structure of an ES member active site (mandelate racemase) is shown above. The residue positions conserved across members of the ES are shown in blue stick representation and a ligand is shown in red surface representation; a metal ion is shown as the green sphere.

by members of the ES [3] include:

EC Number	Family
4.2.1.11	enolase
4.2.1.40	glucarate dehydratase (GlucD)
4.2.1.113	*o*-succinylbenzoate synthase (OSBS)
4.3.1.2	methylaspartate ammonia lyase (MAL)
5.1.2.2	mandelate racemase (MR)
5.5.1.1	muconate lactonizing enzyme (MLE)

As shown above, the ES contains families of enzymes with highly similar enzymatic functions (e.g., 4.2.1.11 and 4.2.1.40) and families that differ in function at all 4 tiers of the EC classification (e.g., 4.2.1.11 and 5.1.2.2). Because EC numbers provide an established set of protein function classification labels, they are used throughout this thesis as a convenient means of assigning proteins to discrete classes. In Ch. 5, EC classes are used to define the set of proteins constituting a single family of functionally related enzymes.

Chapter 3

An Unsupervised Learning Approach: The FASST Framework

3.1 Motivation

As the amount of available protein structure data continues to increase, identifying biologically meaningful structural trends becomes increasingly difficult using conventional approaches such as structure superposition and pairwise structure comparison. Identifying the impact that structural variation has upon the function of related protein structures is becoming increasingly important for understanding the structural basis for functional diversification [26].

For instance comparative structural analysis of ligand binding sites both within and across species has played a significant role in the evaluation of lead compounds as well as finding new targets for existing compounds [18, 27, 28] For rigid docking methods, identifying the number of specific structure conformations to test so that the structural variation of the target is sufficiently represented has a significant impact on the success of these meth-

ods [29]. However, identifying the minimum number of unique structural conformations for a protein requires the adoption of comparative structure analysis approaches capable of considering all available structure data simultaneously.

Approaches for cataloging and classifying structures at the domain level, such as Pfam [30], CATH [31] and the CDD [32], have been heavily developed over recent decades and have become critical resources for understanding the structural landscape of the proteome. However, local structure comparison can provide unique insights into the evolution and function of proteins that are inaccessible to fold- and sequence-based analysis approaches [33, 34].

In order to facilitate the use of comparative local structure analysis, The FASST Framework was developed to provide a general framework for the local structure analysis of large protein datasets in a wide variety of application domains. The approach used by The FASST Framework to allow for easy modification and customization to different types of structure analysis is detailed here.

3.2 Problem statement

The unsupervised learning problem that The FASST Framework supports can be succinctly stated as follows:

> *Given a set of protein substructures, identify the unique structural conformations present.*

To implement this analysis, The FASST Framework provides the following interface:

- **Input:** aligned set of protein substructures
- **Output:** a substructure clustering

In the definition above, a set of r residues from a single protein structure constitutes one substructure. In order for a set of n input substructures to be considered "aligned", a residue-residue bijection among all pairs of the n substructures must exist. This pairwise residue-residue bijection defines a one-to-one correspondence between the r residues of substructures `A` and `B`, for all pairs of the n substructures.

Many approaches could be taken to generate a set of aligned substructures for input to The FASST Framework. For example, in Sec. 3.11.1, the same 4-residue substructure (Glu71, Leu75, Asp168, Phe169) is compared across all available structures for p38α. There, the substructure alignment is trivial because the same 4 residues are compared across different structures of the same protein; that is, Glu71 in structure `A` corresponds to the same Glu71 in structure `B`, and likewise for the remaining residue positions.

Another approach to generating an aligned substructure set, depicted in Sec. 3.11.2, is to use the alignment columns of a multiple sequence alignment in order to determine the residue-residue correspondence across all proteins in the sequence alignment. This approach is used throughout this thesis. Additional alignment approaches are enumerated in Sec. 3.5.

The substructure clustering output by FASST can be used to identify a number of traits for the input substructure set. As demonstrated in Sec.3.11.1, the substructure clustering can be used to identify the number of distinct structure conformations for the residues of the substructures. Outlier substructures can also be distinguished in the example in Sec. 3.11.1 by identifying substructures that are distant from all of the major clusters.

The generality of the method implemented by The FASST Framework was achieved by minimizing the number of assumptions made concerning the nature of the input substructures (i.e., no particular alignment method is assumed) and allowing the major components of The FASST Framework to be independently modified or replaced entirely while still re-

taining the functionality of the remainder of the components (see Sec. 3.10.1 for example usage).

3.3 Related work

Seminal work by Holm et al. 1996 [35] on mapping the protein structure universe demonstrated how an all-against-all (global) structure comparison could reveal the high-level relationships among protein structure folds. A number of approaches to comparative structure and binding site analysis have also been developed and demonstrated to successfully identify biologically meaningful trends. Previous work illustrated by Kinjo et al. 2009 using a surface similarity network approach [36] identified a high-level organization of protein ligand binding site similarity among structures spanning the entirety of the (ligated) PDB. The recently developed Protein Surface Classification (PSC) method illustrated an approach similar to that Kinjo et al. by clustering protein binding site surfaces. Due to the diversification of enzymatic function within divergently related families that share a common fold but exhibit a range of enzymatic activity (such as the ES), the goal of PSC is the functional annotation of protein sequences by providing a finer classification than domain ontologies such as CATH [31] and SCOP [37].

Approaches to efficiently searching large structure datasets for local structure similarity to a query binding site have been thoroughly investigated [38, 39, 40, 41, 42, 43, 44, 45, 46, 2]. A major distinguishing factor among these local structure search and comparison methods is the approach used to quantify local structure dissimilarity. However, given the structural variety of protein functional sites, it is difficult to derive a single measure capable of optimally quantifying *meaningful* similarity/dissimilarity of local structural features in every possible context. For example, while the electrostatic surface potential comparison

approach implemented by HTHQuery [43] is capable of identifying specific DNA binding site motifs with high specificity, the same dissimilarity measure may not be appropriate for distinguishing among binding sites that interact primarily through steric and hydrophobic interactions.

Additionally, the problem of aligning proteins via sequence, structure or a combination thereof has been studied for many decades, producing a variety of both general and highly specialized techniques. General global structure alignment methods such as DaliLite [47], CATHEDRAL [48], CE [49], VAST [50], HOMSTRAD [51] and MATT [5] have proven capable of recognizing fold similarity among highly divergent protein sequences. Sequence-based approaches include HMMER [52], PROSITE [53], CLUSTALW [54] and MUSCLE [55], to name a few. Local structure alignment approaches, such as those used for local search and comparison discussed in the previous paragraph provide an additional approach to identifying corresponding features among proteins in cases where global structure- and sequence-based are incapable [33, 34].

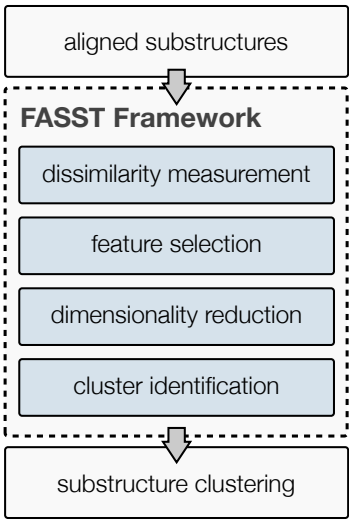
By providing a general framework within which to relate alignment methods with local structural comparison methods mentioned above, The FASST Framework provides a unified approach to combining these approaches to facilitate the large-scale analysis of local structural features as detailed below.

3.4 Method overview

Given the sheer quantity and variety of both alignment and comparison methods that have been developed for both general and highly specific biological applications, The FASST Framework was designed to be agnostic to both the alignment and dissimilarity measure used to compare proteins, allowing the user to customize the analysis to suit the particular

set of proteins being studied.

In order to compute substructure clusterings, FASST first computes the pairwise dissimilarity measure among pairs of the aligned substructures. Next, one feature vector per substructure is constructed from the pairwise comparisons of that substructure with the other aligned substructures. Because the feature vectors representing each substructure are typically of very high dimension, dimensionality reduction is used to compute a low-dimensional approximation of each vector. Finally, the dimensionality reduced feature vectors are clustered into sub-groups. Each of these steps is detailed in sections 3.6-3.9.



3.4.1 Interpreting clustered substructures

In order to build intuition for interpreting the substructure clusterings computed by The FASST Framework, a set of aligned substructures with the corresponding clustering is shown in Fig. 3.1. In this case, the comparison set is a collection of 5-residue enzymatic site substructures, 83 sites in all, that were aligned by LabelHash [2] as previously described in [1].

Each point in the scatter plot (Fig. 3.1**(b)**) is a feature vector that represents a single substructure. The feature vectors are colored by their cluster assignment as determined by FASST. The colored marker labels are present only to illustrate a single example PDB ID for one of the protein substructures in each cluster.

If substructure A is nearby substructure B in the clustering, then these two substruc-

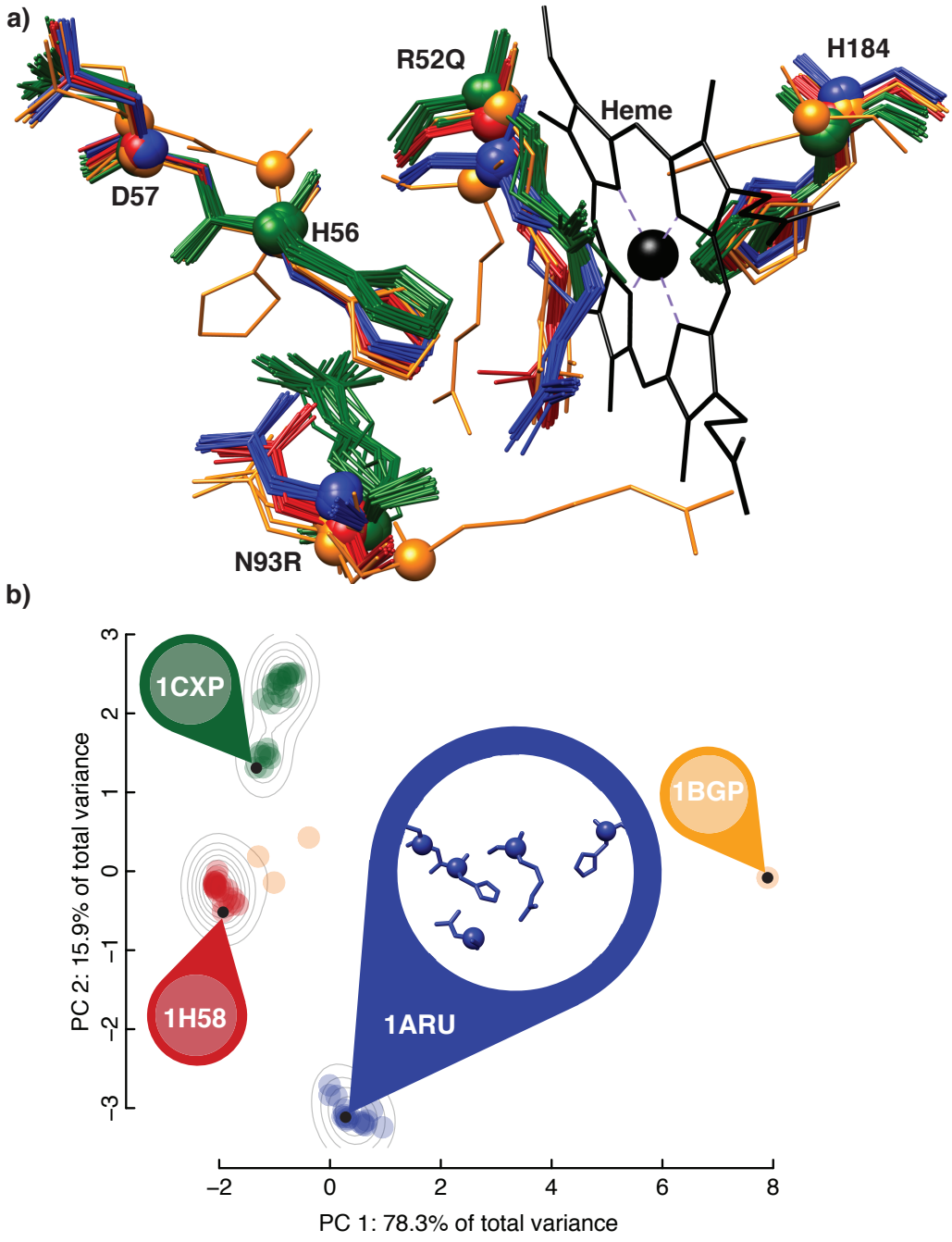


Figure 3.1 : **Substructure clustering.** A set of aligned substructures **(a)** and the corresponding substructure clustering **(b)** as computed by FASST is shown above. As can be seen, 4 clusters have been identified and are denoted by color. Each 5-residue substructure in **(a)** corresponds to a single point (feature vector) in **(b)**. Adapted from [1].

tures are similar, with respect to the dissimilarity measure used, relative to the remaining substructures in the comparison set. Conversely, two substructures that are far apart in the visualization are dissimilar.

In the structural alignment shown in Fig. 3.1**(a)**, all 83 substructures are shown. Each is colored according the cluster assignment identified by FASST. Note that substructures with the same color are well-aligned with one another and are distinguishable from the substructures of a different color. The gold-colored outlier (labeled 1BGP in scatter plot) corresponds to the gold-colored outlier in the full substructure alignment.

3.5 Defining aligned substructures

Because the choice of an alignment approach can be specific to the type, quantity and size of the substructures being analyzed, the generalized The FASST Framework introduced here provides a general approach that is *alignment agnostic*.

In particular, The FASST Framework has previously been combined with several different alignment approaches:

- Local structure-based alignment (LabelHash)
- Global structure-based alignment (CE)
- Progressive sequence-based alignment (CLUSTALW)
- Profile-HMM sequence-based alignment (HMMER)

Choice of an alignment approach depends on many factors including:

- Degree of sequence conservation
- Fold similarity

- Function of the selected residues (e.g., binding site)

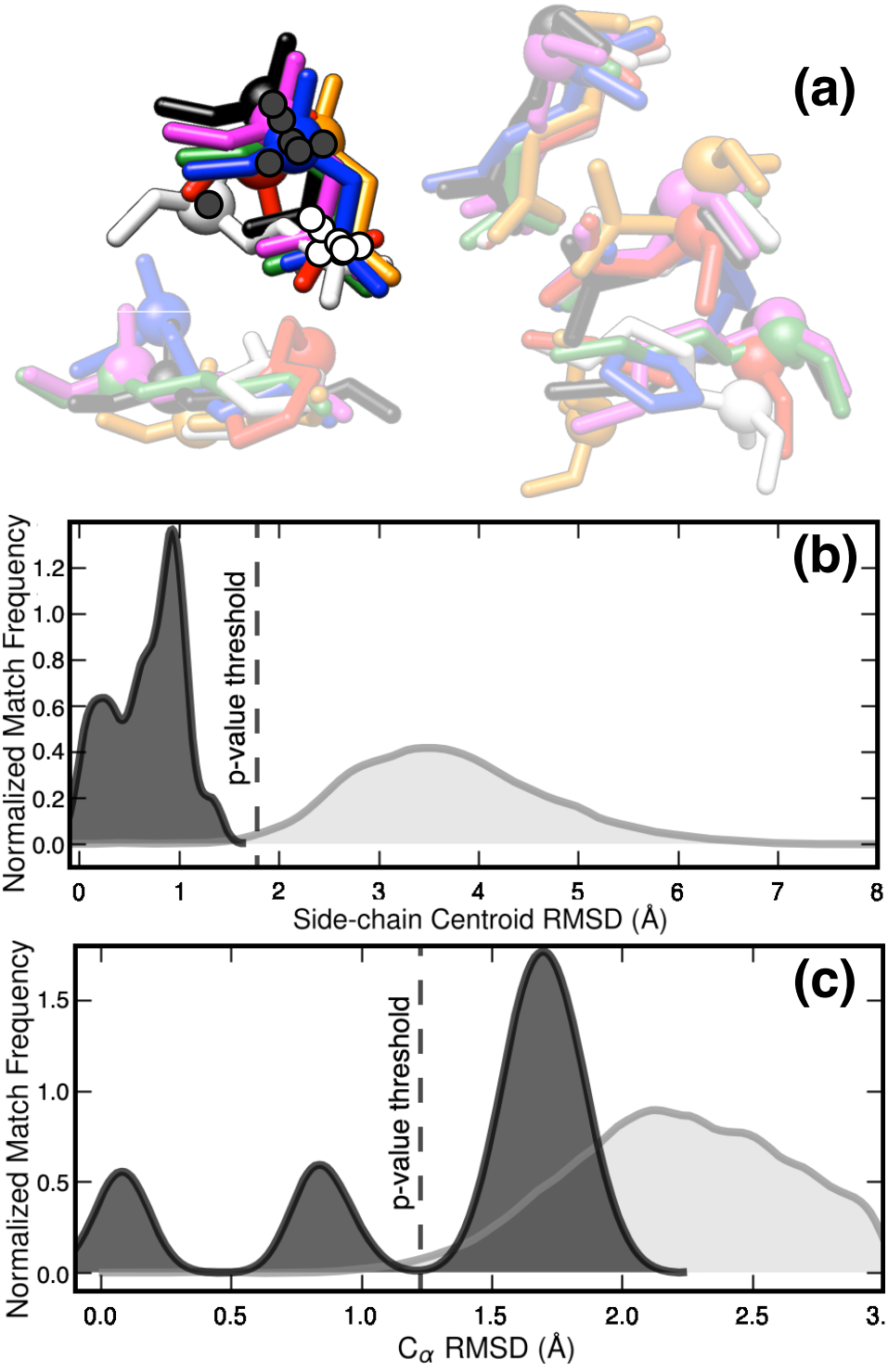
The results presented in this chapter all utilize the HMMER-based alignments provided by Pfam (version 25.0) [30].

3.6 Computing pairwise substructure dissimilarities

The ability to identify functionally similar features among proteins depends directly upon the dissimilarity measure used for quantifying the similarity of a pair of proteins. As a case study in how a dissimilarity measure can drastically effect the ability of computational approaches to identify functional similarity, the proteins of the ES are examined here again. As can be seen in Fig. 3.2**(a)**, ES structures have higher C_α RMSD variability than side chain RMSD variability when examining the 5-residue active site substructure shown in Fig. 3.2. As shown in Fig. 3.2**(b)**-**(c)**, C_α RMSD is not as effective as side chain RMSD in distinguishing ES structures from non-ES structures.

Because side chain positions of the ES motif are more heavily conserved than the C_α positions, side chain RMSD was shown to be a more powerful similarity measure for identifying substructures that share the same ES partial reaction step [2]. However, in cases where side chain positions are less constrained and therefore tend to have a higher variance (e.g., flexible loop regions, protein-protein interfaces), dissimilarity measures other than side chain RMSD may be more appropriate.

In order to facilitate the wide-applicability of The FASST Framework to many different types of substructure analysis, The FASST Framework is not constrained to using any particular dissimilarity measure. As detailed in Sec. 3.6, many different dissimilarity measures can be freely substituted to suit the particular application.


(a)
(b)
Normalized Match Frequency
p-value threshold
Side-chain Centroid RMSD (Å)
(c)
Normalized Match Frequency
p-value threshold
Cα RMSD (Å)

3.6.1 Pairwise dissimilarity measures

Several alternative approaches to computing the structural similarity of pair of substructures are examined and compared here to build intuition for how the choice of a dissimilarity measure for a particular problem will affect the resulting feature vectors computed by The FASST Framework. As shown in Fig. 3.3, The FASST Framework was used to compute feature vectors for all structures of the PFAM:PKINASE family using pairwise comparisons with several different dissimilarity measures. The structure shown in Fig. 3.3**(e)** corresponds to a p38 kinase structure with bound imatinib molecule (PDB:3HEC); the 5-residue substructure shown corresponds to the C-helix (E71, L75), gatekeeper (D106) and the DFG-motif (D168, F169) residues. The feature vectors computed by The FASST Framework using C_α RMSD, side chain RMSD, BLOSUM-RMSD and TRAIT-RMSD are shown in Fig. 3.3**(a)** through **(d)**, respectively. The procedure for calculating each of these dissimilarity measures is detailed in the following sections.

Figure 3.2 *(preceding page)*: **Effect of dissimilarity measure on identification of enolase superfamily members** The ability of LabelHash [2] to distinguish proteins in the ES from non-ES structures is greatly dependent upon the dissimilarity measure used for scoring structural similarity during the structure search. The substructure responsible for the conserved partial reaction among members of the ES consists of the 5 residues from which the ES motif used here is derived [3]. This 5-residue substructure is shown in **(a)** for 7 different ES members. One of the 5 residues is highlighted to illustrate the deviations among the C_α and side chain centroid positions that are marked with dark and light gray circles, respectively. In the case of ES member proteins, the C_α positions have greater variability than the side chain positions. In **(b)** and **(c)**, the structure matches identified using side chain centroid and C_α RMSD, respectively, are compared. The *x*-axis in each plot above denotes the RMSD of the ES motif to identified matches and the *y*-axis denotes the normalized number of structures that matched the motif at a particular RMSD. Dark gray and light gray denote matches to ES and non-ES structures, respectively. The dashed lined corresponds to the LRMSD distance threshold for matches having statistically significant similarity as identified by the LabelHash statistical model. As can be seen above, side chain RMSD clearly distinguishes ES and non-ES structures while C_α RMSD does not.

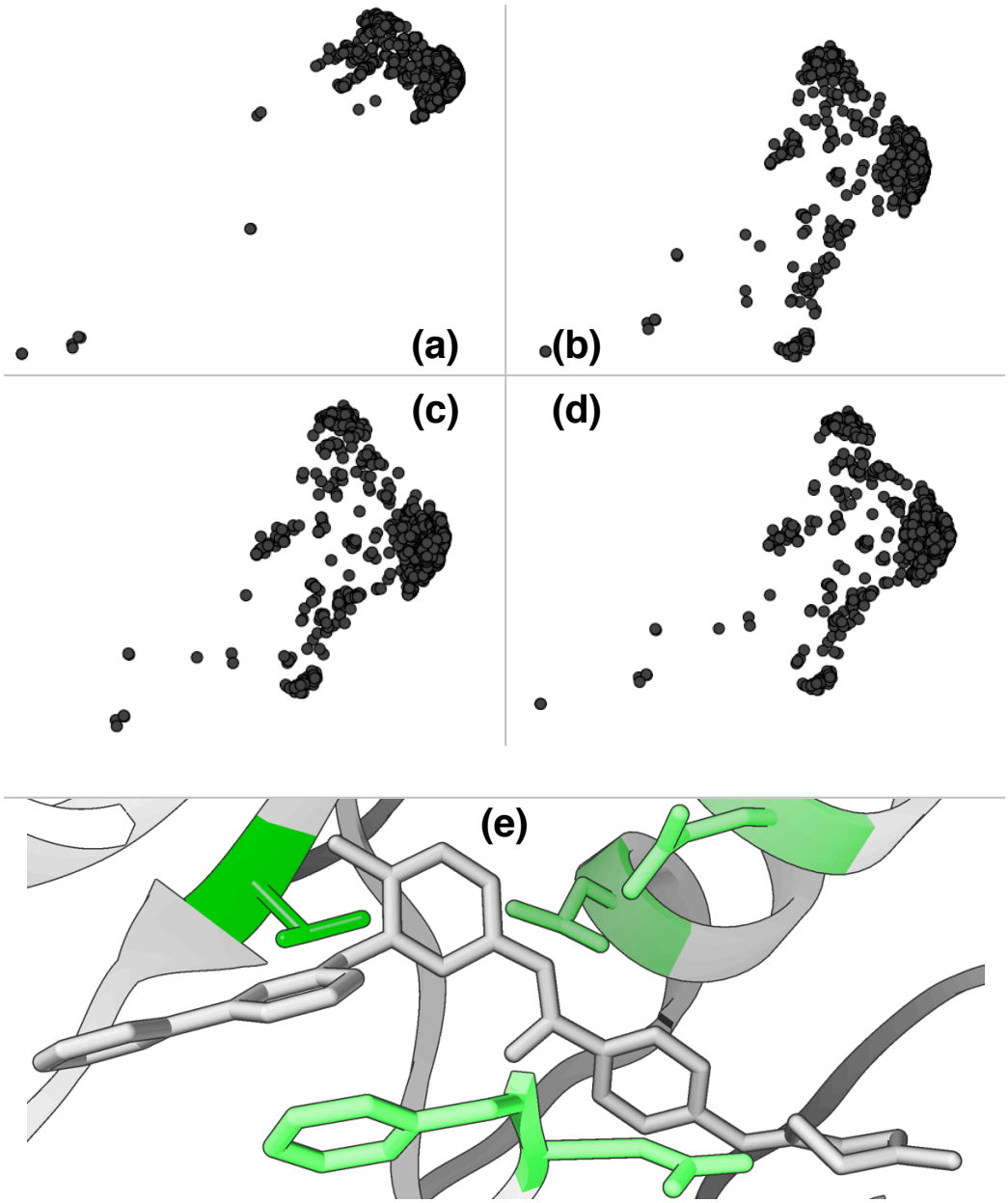


Figure 3.3 : **Effect of dissimilarity measure on the feature vectors computed by The FASST Framework** . **(a)** C_α RMSD, **(b)** side chain RMSD, **(c)** BLOSUM-RMSD, **(d)** TRAIT-RMSD, **(e)** the 5-residue kinase substructure being compared among structures of the PFAM:PKINASE family in **(a)** through **(d)** above. Comparison of the feature vectors between **(a)** and **(b)** above illustrates that side chain RMSD is more variable than C_α RMSD among the PFAM:PKINASE family. The combination of side chain RMSD with a side chain residue dissimilarity measure has a more subtle affect on the feature vector distribution as shown by comparing **(b)** to **(c)** and **(d)** above.

Geometry-only dissimilarity measures

The Root Mean Square Deviation (RMSD) between a pair of aligned protein structures is one frequently used measure of structural similarity and is calculated as:

$$\text{RMSD} = \sqrt{\sum_{i=1}^{m} \frac{\Delta x^2 + \Delta y^2 + \Delta z^2}{m}} \tag{3.1}$$

for a set of m aligned points. In the case of C_α RMSD, the aligned points are the C_α atoms of each residue. As discussed in Sec. 3.5, both global and local approaches to calculating pairwise alignments exist, such as CE (global) or LabelHash (local), and C_α RMSD can be used to quantify the similarity of a pairwise alignment in both cases. For side chain RMSD, the aligned points are instead the coordinates of the side chain centroids for each residue.

Geometry-augmented dissimilarity measures

Geometry-only dissimilarity measures, such as C_α RMSD, will score a pair of substructures having highly similar backbone geometry as being very similar, regardless of possible mutations or residue substitutions that may be present in one of the substructures. For applications where local fold similarity is the quantity of interest, disregarding residue substitutions and only considering backbone geometry may be appropriate. However, when comparing the binding site substructures among a homologous family of proteins, such as the protein kinases, that share a large degree of fold similarity within the binding site region, geometry-only dissimilarity measures such as C_α RMSD may have very little variance and fail to capture the important substitutions of functionally important residue positions.

A simple approach to incorporate residue substitution penalties into a geometric dis-

similarity measure is to simply sum the respective dissimilarity measures:

$$d = d_{\text{side chain centroid}} + d_{\text{residue substitution}} \tag{3.2}$$

Given the above approach for combining a geometric and residue substitution penalty, it is still necessary to define how residue substitution is quantified between a pair of substructures, where each substructure is composed of multiple residues.

Canonical residue substitution matrices, such as BLOSUM62 [56] or PAM120 [57], that are frequently used during sequence alignment, provide a well-understood basis for quantifying the likelihood of a residue substitution at a particular sequence position. These substitution matrices are typically provided as a 20×20 matrix that defines the penalty for substituting any residue `A` for any other residue `B`. The major distinction between the different types of substitution matrices is the algorithm used for computing the likelihood of each substitution.

The following procedure was used to convert the canonical residue substitution matrices that are typically represented as a matrix of residue-residue scores with larger values indicating more likely substitutions. Given matrix M having similarity score values $s_{i,j} \in [min(M), max(M)]$, the following residue-residue dissimilarity matrix is created:

$$M' = \frac{-M + max(M)}{max(M) - min(M)} \tag{3.3}$$

The resulting dissimilarity matrix M' contains elements $d_{i,j} \in [0,1]$ with larger values indicating greater dissimilarity. The dissimilarity of a pair of substructures, each containing m residues, is then calculated as:

$$d_{\text{residue substitution}}(s_1, s_2) = \frac{\sum_{i=1}^{m} M'[s_1^i, s_2^i]}{m} \tag{3.4}$$

where $M'[i,j]$ is the dissimilarity of residues i and j as calculated above and s_1^i is the ith residue of substructure s_1.

An alternative approach to measuring residue-residue similarity that has been demonstrated to be effective for identifying functional similarity among small molecule binding sites of the protein kinases is the pharmacophore trait measure introduced by [58]. The dissimilarity between a pair of substructures is quantified by a combination of chemical feature traits, such as aliphaticity, number of hydrogen bond donors and acceptors, aromaticity and hydrophobicity. Combining the dissimilarity scores for each of these traits with a geometric dissimilarity measure, such as side chain RMSD, yields the following:

$$\begin{aligned} d(s_1,s_2) = {} & d_{\text{side chain centroid}}(s_1,s_2) + d_{\text{size}}(s_1,s_2) + \\ & d_{\text{aliphaticity}}(s_1,s_2) + d_{\text{aromaticity}}(s_1,s_2) + \\ & d_{\text{hydrophobicity}}(s_1,s_2) + d_{\text{hbond acceptor}}(s_1,s_2) + \\ & d_{\text{hbond donor}}(s_1,s_2). \end{aligned}$$

This dissimilarity measure is referred to hereafter as the TRAIT-RMSD and is used throughout Ch. 4–6 for all pairwise binding site substructure comparisons.

3.6.2 Comparison of dissimilarity measures

As shown in Fig. 3.3, the biggest affect on the feature vectors computed by The FASST Framework in the case of the protein kinase substructure shown in Fig. 3.3**(e)** comes from using side chain RMSD rather than C_α RMSD. This result is not particularly surprising considering the backbone positions of the residues in **(e)** are more constrained than the side chain positions. The addition of the residue substitution dissimilarity measure as shown

in **(c)** and **(d)** has a much less dramatic effect on the number of identifiable feature vector clusters.

3.7 Computing feature vectors

In order to compute the per-substructure feature vectors for a set of substructures, FASST computes a dissimilarity matrix for the substructures. Given n substructures, the full all-vs-all dissimilarity matrix would be comprised of $n(n-1)/2$ unique comparisons (due to dissimilarity measure symmetry). Cell $(i,\ j)$ within the dissimilarity matrix corresponds to $d(s_i, s_j)$ where d is the dissimilarity function described in the Section 3.6 and s_i and s_j are both protein substructures. Each row of the dissimilarity matrix is considered a feature vector, where row i of the dissimilarity matrix is a feature vector representing the structural dissimilarities of substructure s_i to the remainder of the substructures in the matrix.

However, due to the fact that the dissimilarity matrix computation grows as $O(n^2)$ with the number of substructures analyzed, the feature vector computation step can become prohibitively expensive to compute for very large numbers ($>$1000) of substructures. To address this issue, two approaches to approximating the full dissimilarity matrix while only computing a fraction of the $O(n^2)$ comparisons are introduced below.

3.7.1 Randomized landmark selection

The most computationally expensive step of FASST is the calculation of the $n \times n$ dissimilarity matrix. Because the clustering step requires only the dimensionality-reduced form of the feature vectors, it is not strictly necessary to compute the full $n \times n$ dissimilarity matrix. Because of the large degree of structural overrepresentation for many proteins, it is possible to randomly sample a subset of the dissimilarity matrix columns and then apply the

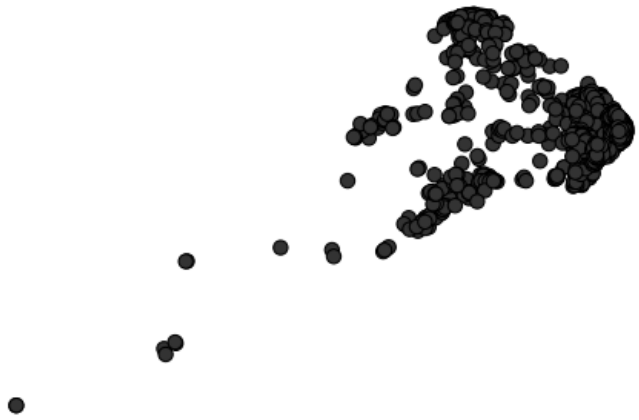


Figure 3.4 : **Dimensionality reduced feature vectors computed from the full $n \times n$ dissimilarity matrix.** The feature vectors computed from the full $n \times n$ dissimilarity matrix, for the 5-residue substructure shown in Fig. 3.3**(e)**, across all of the PFAM:PKINASE family proteins, are shown above.

same dimensionality reduction and clustering steps (Section 3.8 and 3.9) while preserving the same cluster membership. This is in fact a common technique in dimensionality reduction [59]. By randomly sampling columns from the dissimilarity matrix, the computation is reduced from $O(n^2)$ to $O(kn)$ where $k \ll n$ typically, allowing The FASST Framework to scale easily as additional structures become available.

The columns of the dissimilarity matrix that will be sampled are referred to hereafter as "landmarks." While dimensionality reduction is detailed in the following Sec. 3.8, it is important to mention that the comparisons made here in assessing the landmark selection approaches are based upon comparing the dimensionality reduced versions of the feature vectors, which for consistency have been arbitrarily made to be 2-dimensional.

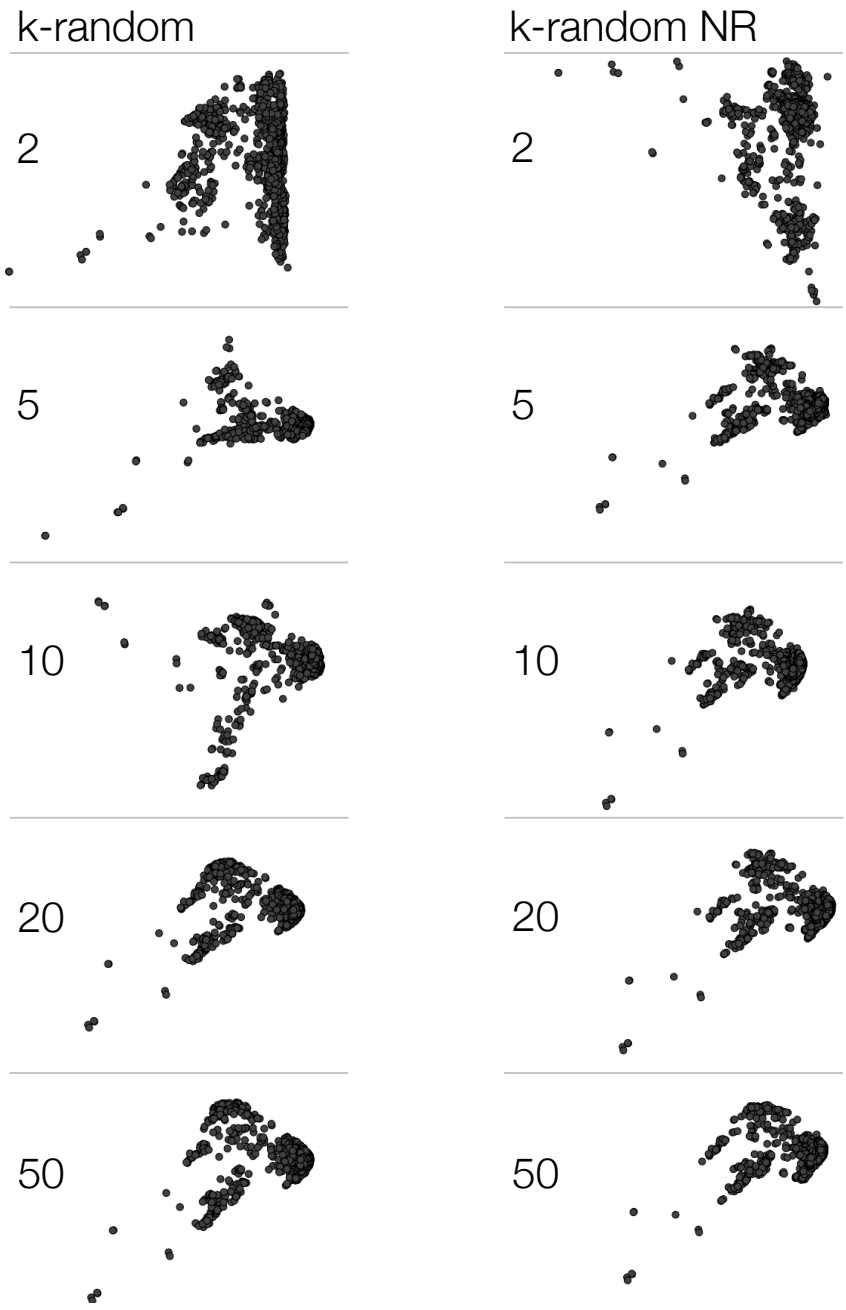


Figure 3.5 : **Dimensionality reduced feature vectors computed from a $k \times n$ dissimilarity matrix subset**. The feature vectors computed from a $k \times n$ subset of the full $n \times n$ dissimilarity matrix for the 5-residue substructure shown in Fig. 3.3**(e)**, across all PFAM:PKINASE family proteins, are shown above. The two different approaches to landmark sampling, uniform random and non-redundant random, are compared for $k \in \{2, 5, 10, 20, 50\}$. As can be seen above, the k-random NR approach produces a clustering very near that of the full dissimilarity matrix (compare to Fig. 3.4) by $k = 5$, as opposed to $k = 20$ for the uniform random approach.

The simplest landmark sampling approach is to select k landmarks uniformly at random. Given this approach, the value selected for k should ideally be as small as possible to sufficiently approximate the feature vectors of the full dissimilarity matrix (after dimensionality reduction). The feature vectors computed using k uniformly random landmarks for $k \in \{2,5,10,20,50\}$ are shown in Fig. 3.5 and can be compared with the feature vectors for the full dissimilarity matrix shown in Fig. 3.4. For $k \in \{20,50\}$, the resulting feature vectors shown in Fig. 3.5 are a close approximation of the full feature vectors. However, for $k \in \{2,5,10\}$ significant deviations are apparent.

Because the landmarks selection is non-deterministic, repeated trials will select different sets of landmarks. By repeating the feature vector computation with many different random landmark selections while keeping k constant, it was noted that some random selections produced very close approximations even at the $k = 5$ level, but the variance in quality was found to be high. Because the feature vector computation step is critical to the following cluster identification step, a value for k that was able to reliably reproduce a close approximation of the full dissimilarity matrix is necessary. Further inspection of the landmarks selected in the degenerate cases revealed that the primary cause of the high approximation variance was the extreme overrepresentation of a small number of protein sequences. These overrepresented sequences were sampled multiple times in the degenerate cases, causing a low amount of diversity among the selected landmarks.

3.7.2 Redundancy-aware landmark selection

In order to address the problem of structurally overrepresented sequences during landmark selection an alternative redundancy-aware approach was developed. By sampling from a non-redundant subset of the protein sequences present among the substructures, the worst-

case behavior of the uniform random selection where all landmarks correspond to a single protein sequence can be avoided completely.

The redundancy-aware landmark sampling approach implemented first partitions a protein family into sequence non-redundant groups at the 50% sequence identity level (NR_{50}-clusters). Then, the highest resolution structure from k randomly selected non-redundant groups is selected as a landmark. For a set of substructures having fewer than k NR_{50}-clusters, each of the NR_{50}-clusters will contribute one representative structure, resulting in $|NR_{50}\text{-clusters}|$ landmarks.

As shown in Fig. 3.5 for $k \in \{2,5,10,20,50\}$, randomly selecting k landmarks from distinct NR_{50}-clusters results in feature vectors that closely approximate the full dissimilarity matrix for smaller values of k when compared to the uniform random selection procedure. To be conservative, $k = 20$ for the redundancy-aware landmarks selection procedure was selected as the default method for The FASST Framework. Note that while randomized landmark selection will inherently select a different random set of landmarks on repeated executions, the process can be made repeatable by seeding the random number generation process with a constant value across trials.

3.8 Dimensionality reduction

In order to reduce the dimensionality of the computed feature vectors, Principal Components Analysis (PCA) [60] is used to transform the raw vectors of the computed dissimilarity matrix (or matrix subset) into a lower dimensional embedding. Only the first two principal components are kept by default and the remainder of the components are truncated, thereby resulting in a 2-dimensional feature vector representation of each of the n protein structures in the dataset (i.e., an $n \times 2$ sub-matrix of the transformed vectors). Previous

work [1] demonstrated that the amount of data variance present in the identified principal components dropped sharply after the first 2 components.

However, a significant structural redundancy problem arises when performing the dimensionality reduction of the feature vectors. By default, PCA weights the importance of all feature vectors equally when computing the low-dimensional embedding that optimally preserves the variance of the original data. Overrepresented protein sequences cause PCA to place unequal emphasis on preserving the variance among structures for the overrepresented sequence rather than equitably across all of the sequences in the dataset. This bias in the embedding effectively hides variation among distinct protein sequences in favor of preserving the variation among the overrepresented sequence.

To remove overrepresentation bias, PCA is first computed with a non-redundant subset of the rows of the dissimilarity matrix; that is, one feature vector is computed for each NR_{50}-clusters. Then PCA is computed for this sequence non-redundant set of feature vectors alone, thereby avoiding the bias induced by structurally overrepresented sequences. Next, the feature vectors are computed for all structures in the dataset and then transformed to a lower dimensional embedding using the PCA transformation matrix computed from the non-redundant subset of the dataset. This approach amounts to computing a binary weighted PCA where all feature vectors have a weight of 0 with the exception of the non-redundant subset of structures that have a weight of 1.

3.9 Clustering feature vectors

Given the dimensionality reduced feature vector representation of the substructures, these vectors are then clustered to identify sub-groups that share strong structural similarity. However, the number of clusters to expect is difficult to quantify beforehand and has been

noted in previous work [1] to vary greatly depending on particular substructure being analyzed.

The Gaussian Mixture Model (GMM) clustering method as implemented in the MCLUST package [61] is able to identify both the number of clusters present and the cluster memberships for each of the feature vectors. The GMM approach was demonstrated in previous work [1] to identify clusterings that corresponded to biological properties such as ligation state and fold class.

However, a single-pass GMM clustering approach is not sufficient to identify the number of clusters present. The approach implemented by MCLUST is to iteratively increase the number of multivariate Gaussian distributions present in the mixture model and then evaluate the fitness of a particular k-Gaussian model using the Bayesian Information Criterion (BIC). Because BIC is a regularized approach, given 2 models with equally good fit to the feature vectors, the model having fewer Gaussians is preferred (assuming an equal number of covariance matrix parameters per Gaussian are allowed for both models).

3.10 Design and implementation of The FASST Framework

To facilitate the application of The FASST Framework to a wide variety of structural comparison problems, the framework is designed to have a highly modular architecture with components having only a minimal interface to facilitate reuse and ease of customization. An overview of the components making up The FASST Framework is shown in Fig. 3.6

3.10.1 Components

As shown in Fig. 3.6, The FASST Framework is composed of 5 major types of components that each expose only a minimal interface with one another. The components can in fact

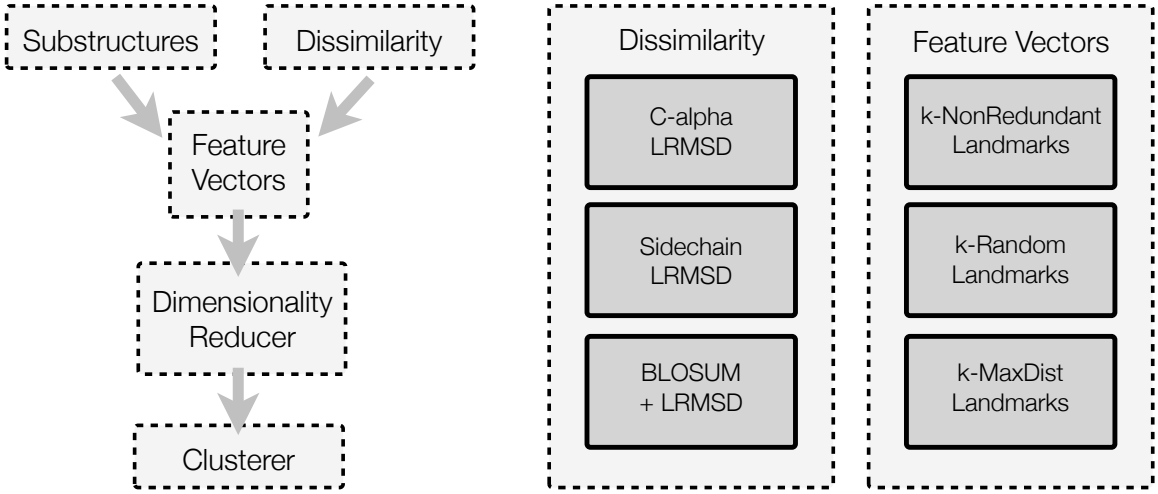


Figure 3.6 : **Design of The FASST Framework** The FASST Framework provides a modular architecture within which components, such as dissimilarity measures can be easily swapped and modified without requiring modifications to other components. A sampling of interchangeable dissimilarity measures and feature vector computation procedures are shown within solid rectangles.

be thought of as a loosely associated set of data transformations that have the specific interfaces described below.

Dissimilarity measures

All dissimilarity measures are assumed to be pairwise comparisons having the function signature: $d(s_1, s_2)$ and return a floating point value $x \geq 0$.

Feature vector computations

A feature vector computation has the function signature: $f(S, d)$ and returns an length-n array of m-dimensional vectors, where S is a set of n substructures and d is a dissimilarity measure. The resulting $n \times m$ feature vector matrix is labeled $\mathbf{V}$ hereafter.

Dimensionality reducers

A dimensionality reducer has the function signature: $r(\mathbf{V})$ and returns a new $n \times m'$ set of feature vectors $\mathbf{V}'$, such that $m' \leq m$.

Clusterers

A clusterer has the function signature: $c(\mathbf{V}')$ and returns a length-n array of integers indicating the cluster ID assigned to each of the n substructures.

All of the variants of each component type discussed above reside in separate modules and have been implemented in the Python programming language. An example of how to construct a custom analysis pipeline using components from The FASST Framework is demonstrated in Listing 1.

```
# select components to use
from fasst.featurevecs import compute_features
from fasst.distances import calpha_lrmsd
from fasst.dimreducers import pca_reduce
from fasst.clusterers import gmm_cluster

# define the set of substructures to analyze
substructures = [s1, s2, ..., sN]

# compute the feature vectors
feature_vectors = compute_features(substructures, calpha_lrmsd)

# reduce the dimensionality
dimreduced_feature_vectors = pca_reduce(feature_vectors)

# cluster the feature vectors
clustering = gmm_cluster(dimreduced_feature_vectors)
```

Listing 1: An example Python session demonstrating how components from The FASST Framework can be combined to construct an application-specific implementation.

In the example shown in Listing 1, a set of substructures that have been fully defined by the user are paired with the `calpha_lrmsd` dissimilarity measure and used to compute a set of feature vectors. The dimensionality of the vectors is next reduced by applying the `pca_reduce` transform. Finally, the clustering of the original substructures is determined by applying the `gmm_cluster` to the dimensionality-reduced feature vectors.

3.11 Interactive visualization and analysis with FASST Live

FASST Live builds on top of The FASST Framework by providing multiple automated approaches to generating a set of substructures to analyze and providing a web-based visualization and exploration interface to the resulting clustering. Each of the pipelines implemented by FASST Live provides a different form of analysis by modifying how the substructures are generated from the user query and are explained in detail below. In addition to the automated pipelines the user can submit a pre-computed substructure alignment to allow greater flexibility and facilitate the application of The FASST Framework to a wide variety of alignment approaches.

3.11.1 Single-sequence, multi-structure pipeline

As illustrated in Fig. 1.1, many protein sequences within the PDB contain a multitude of available structures. For example, at the time of this writing p38 alone has 160 available crystallographic structures in the PDB which have been derived under different crystallization conditions, mutations and bound ligands. In order to explore the structural variety present among all available p38 structures (or any other sequence of interest), FASST Live provides a single-sequence, multi-structure pipeline.

As an example, the single-sequence, multi-structure feature vectors computed for a 4-

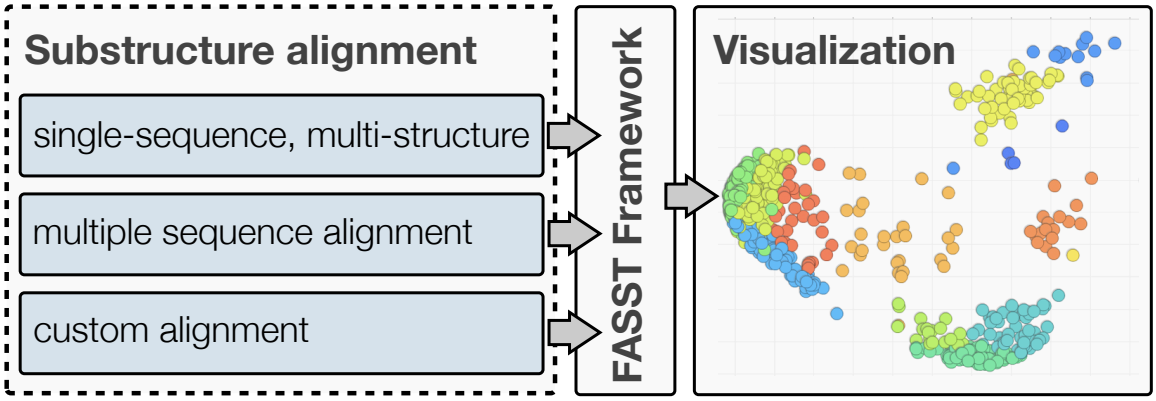


Figure 3.7 : **FASST Live** Each of the pipelines within FASST Live implements a different approach to gathering substructures for analysis with The FASST Framework. Additionally FASST Live provides a rich graphical visualization of the clustered substructures that allows the user to explore high-level structural trends at multiple levels of detail.

residue p38 substructure (residues 71, 75 (C-helix), 168 and 169 (DFG motif)) is shown in Fig. 3.8. The set of substructures compared is obtained by automatically selecting all protein structures in the PDB that have the same UniProtKB sequence ID as the reference substructure (PDB:3HEC). An annotated screenshot of the interactive FASST Live visualization is shown also in Fig. 3.8.

Immediately upon examination of the feature vectors in Fig. 3.8 it is apparent that 3 major substructure conformations exist for p38 (blue, green and yellow clusters) and also many outlier substructures (cyan and orange clusters). Further inspection of substructures in the green cluster (e.g., PDB:3HEC) reveals they share the DFG-out conformation and those in the blue cluster (e.g., PDB:3DT1) reveals that they correspond to DFG-in conformations. Finally, inspection of substructures from the yellow cluster (e.g., PDB:3IW7) reveals a 3rd conformation that is unlike both the DFG-in and DFG-out conformations as shown in Fig. 3.8**(e)**. The high-level trends among the p38 substructures revealed the distinct conformations of the DFG motif and identified a 3rd distinct type of conformation.

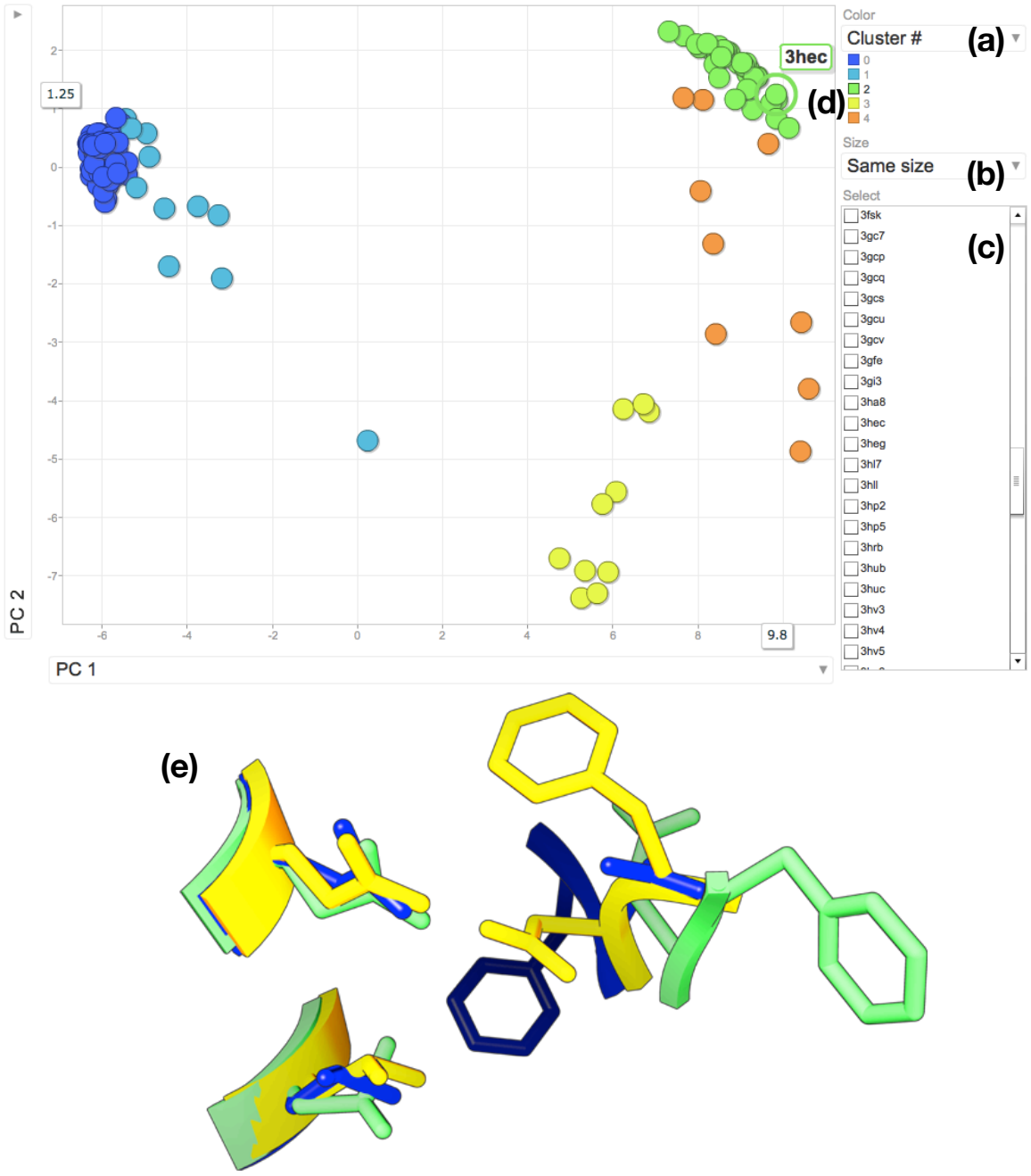


Figure 3.8 : **FASST Live interactive data visualization** In order to gain a deeper understanding into the clustering computed by FASST Live the color **(a)** and size **(b)** applied to each feature vector can be modified. The location of feature vectors for specific substructures can be highlighted by hovering **(d)** or checking the PDB ID in the sorted list **(c)**. The visualization is implemented using a motion chart from the Google Chart Tools API. One substructure from each of the major clusters is shown in **(e)**. The green feature vectors in the chart correspond to DFG-out conformations while the blue correspond to DFG-ing conformations. The yellow feature vectors correspond to conformations that are neither DFG-in nor DFG-out as demonstrated in **(e)**.

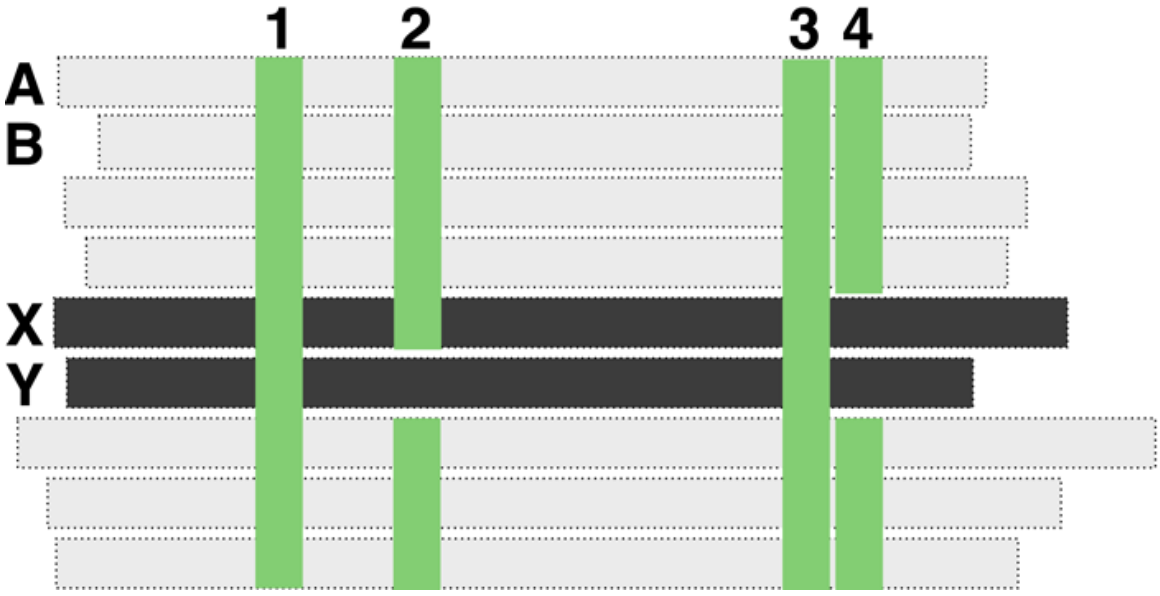


Figure 3.9 : **Procedure for aligned substructure selection from MSA.** The horizontal bars denote aligned sequences within the MSA. The vertical bars denote columns within the MSA that correspond to the selected substructure residue positions. The gaps in the vertical bars represent missing residues within sequences `X` and `Y`. Substructures for sequences `X` and `Y` are excluded because both lack one or more residues at the selected comparison positions within the alignment.

3.11.2 Pfam-based MSA pipeline

Using a multiple sequence alignment (MSA) from the curated Pfam [30] protein family database, corresponding substructures can be selected for all proteins in the MSA. The only additional information needed from the user in order to run the Pfam alignment pipeline (versus the single-sequence pipeline) is the Pfam ID for the protein family of interest.

In order to identify a substructure for sequence `B` given a reference substructure for sequence `A`, MSA columns for the residues specified for sequence `A` are identified then used to identifying the corresponding residue positions for sequence `B`. For example, as shown in 3.9, nine protein sequences of an MSA are shown as horizontal bars and four residue positions are highlighted as green vertical bars.

An MSA alignment, like that illustrated in Fig. 3.9, provides a means of mapping residue positions among proteins. In the case of our previous p38 example, the reference substructure used corresponding residue numbers 71, 75, 168, and 169 of PDB structure PDB:3HEC. In the MSA illustration above, the 4 residues of our reference substructure are denoted by green vertical bars and numbered 1 through 4. However, not every protein sequence in the MSA will have a corresponding residue to the reference substructure, as illustrated by gaps in the green vertical bars. These gapped sequences are excluded from the substructures generated (sequences labeled `X` and `Y` in Fig. 3.9). All other protein sequences in the MSA will be included in the comparison set generated.

3.11.3 User-defined alignment pipeline

Finally, the user can submit a fully defined set of aligned substructures using the custom pipeline option, allowing the user to use any alignment method available to them as input to FASST Live. Such an alignment may be computed by other algorithms for structure (e.g., CE [49], SMAP [62], LabelHash [2]) or sequence alignment (e.g., MUSCLE [55] or CLUSTALW [54]).

The only constraint imposed is that all substructures have an identical numbers of residues specified. This constraint is due to the fact that the desired residue-residue correspondence (between substructures) is derived from the order of residue numbers. The required format for user-defined queries is comma-separated values (CSV) with the following order:

```
pdbid, chainid, r0, r1, r2, ..., rk
```

For example, the following comparison set definition specifies 6 different protein sub-

structures, each with 5 residues:

```
2gb8, A, 48,47,51,180,178
2jti, A, 48,47,51,180,178
3fm6, A, 43,42,46,174,172
3fmu, A, 43,42,46,174,172
2wd4, A, 38,37,41,168,166
1jdr, A, 48,47,51,180,178
```

In summary, the available analysis pipelines and interactive visualization provided by FASST Live provides a web-based service for quickly computing substructure clusterings with The FASST Framework without requiring the user to write code or install any additional software.

3.12 Conclusion

The FASST Framework provides the first general toolkit for constructing custom analysis pipelines for large-scale, local structure analysis. Importantly, the methods describe allow for the full incorporation of all available structure data without need for sequence redundancy filtering and pruning. As demonstrated in Sec. 3.11.1, even the analysis of structure data for *a single* protein sequence can identify unexpected structural trends, such as a 3rd heavily populated conformation of the `DFG` motif, that would not be possible to identify with only a single representative structure.

The substructure clusterings identified by FASST have been demonstrated both here and previously [1] to be capable of identifying biologically meaningful structure trends among both proteins sharing a common fold and proteins sharing *only* local structural similarity at a handful of binding site residue positions.

The dissemination of the general analysis framework introduced here as well as the availability of the interactive FASST Live analysis platform will facilitate the incorporation of high-level local structure analysis among the growing protein structure databases. Finally, the modularity and open source nature of The FASST Framework will allow for further community-driven development and customization for application to a wider variety of biological applications.

Chapter 4

A Supervised Learning Approach: CCORPS

4.1 Motivation

In addition to the increasingly large number of available protein structures, a wide spectrum of large, well-curated annotation databases now exist [63, 64, 25, 30, 65, 66, 67, 68, 69, 70, 71, 72]. By mapping the sequence- and structure-specific annotations within these databases to large structure datasets, a greater understanding of the source of structure variation can be identified as demonstrated in previous work [1].

The FASST Framework approach discussed in the previous chapter provides a critical foundation for constructing redundancy-aware unsupervised learning methods capable of identifying high-level trends among the structural features of large protein datasets. By the unsupervised nature of The FASST Framework, the method does not take into account any additional biological knowledge beyond the sequence and structure of the proteins analyzed. This approach is particularly useful when performing an unbiased analysis of struc-

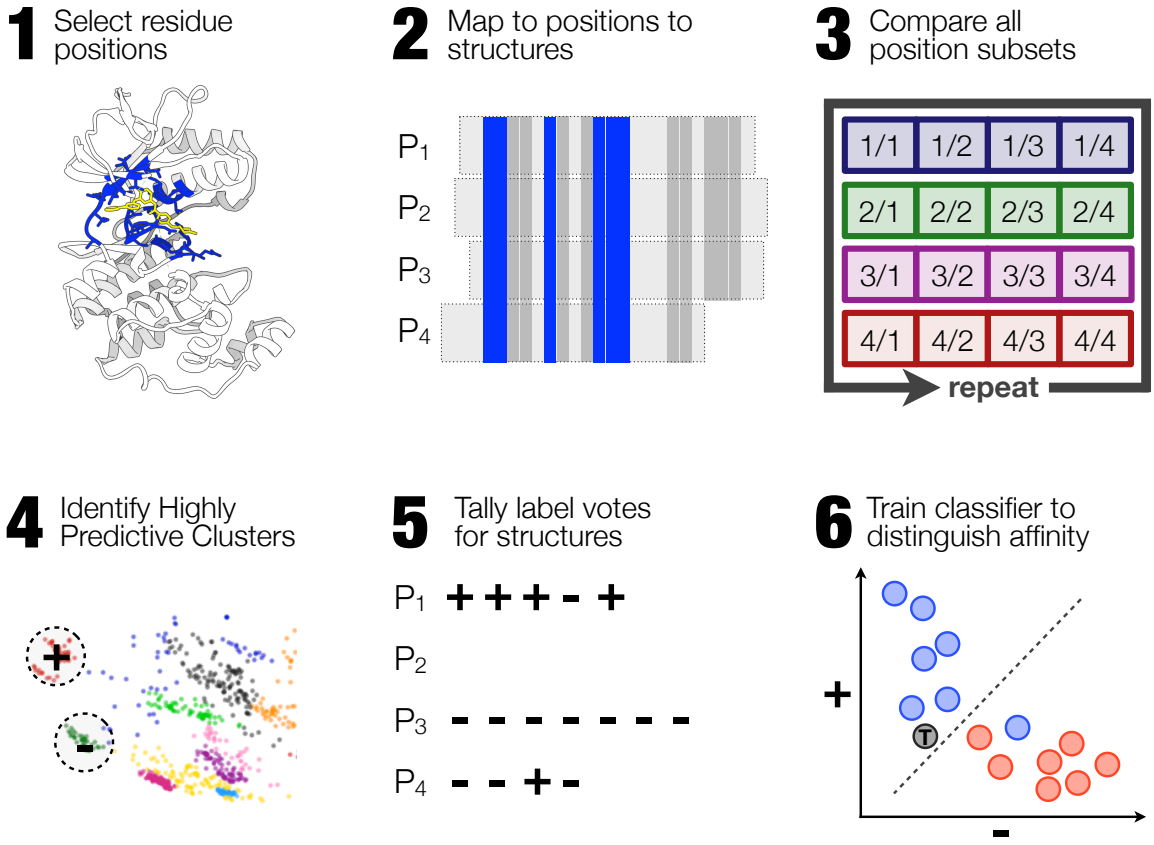


Figure 4.1 : **CCORPS overview**.

ture variation for the purposes of identifying unique structure conformations or structure outliers as discussed in Ch. 3.

However, in cases where annotations are available for the proteins being analyzed, a new formulation of the structure analysis problem can be made. The FASST Framework required knowledge of the specific substructure residue positions to analyze, but identifying the residue positions responsible for structure variation that correlate with a particular set of annotation labels *a priori* is highly non-trivial. The Combinatorial Clustering Of Residue Position Subsets (CCORPS) method, introduced in this chapter, addresses this problem by analyzing all possible residue position subsets in order to identify the structure features that correlate with a given set of annotation labels. The CCORPS method then uses the identified

structure features to predict the annotation labels of other structures lacking annotation, as detailed in the following sections.

4.2 Problem statement

The supervised learning problem that CCORPS addresses can be stated as follows:

> *Find the structural features among the set of proteins that are correlated with a particular set of annotation labels.*

To implement this analysis, CCORPS provides the following interface:

- **Input:** aligned set of protein substructures
- **Input:** set of annotation labels
- **Output:** set of predicted annotation labels

For example, the per-substructure annotation labels may be derived from a wide range of sources [63, 64, 25, 30, 65, 66, 67, 68, 69, 70, 71, 72]. In particular, the applications of CCORPS presented in Ch. 5 and 6 incorporate EC class annotations and inhibitor binding affinity annotations, respectively.

Furthermore, the annotation labels provided to CCORPS do not have to be available for all substructures in the input dataset. Because CCORPS is *semi-supervised* (see Sec. 4.5), annotation labels can be provided for only the subset of substructures where they are available. If, for example, n substructures are provided as input to CCORPS along with m annotation labels (one or zero labels per substructure), where $m < n$, then the output of CCORPS is the $(n-m)$ annotation label predictions for substructures lacking annotation.

In order to assess prediction performance of CCORPS, cross-fold validation experiments (Sec. 4.7) were performed, as discussed in Sec. 5.5 and Sec. 6.5. During cross-fold validation, a subset of the known annotation labels are masked, CCORPS is trained using the remaining unmasked labels and finally CCORPS makes predictions for the masked labels. This process is repeated for each of the validation folds as detailed in Sec. 4.7.

The aligned substructures can be derived using any of the approaches suggested in Sec. 3.2 or Sec. 3.5.

4.3 Related work

In addition to the predicted annotation labels that CCORPS provides, the predictive structural features identified by CCORPS in computing the annotation label predictions are interesting and useful in and of themselves. In some instances, the predictive structure features identified by CCORPS may be related to Specificity Determining residue Positions (SDPs).

Existing structure-based supervised learning approaches such as the FEATURE framework [73] have been demonstrated for the learning of structural features representative of many types of functional sites (e.g., enzymatic, ion-binding). Supervised learning methods such as FEATURE have proven capable of identifying structural indicators of specific functional sites by comparing the physicochemical properties of local micro-environments within enzymes [74, 73, 75]. FEATURE demonstrates the ability of structure-based learning approaches to identify structural trends within large structure datasets.

Incorporating structure data provides an additional dimension to the residue substitution pattern data that exists within MSAs. In addition to the chemical similarity that can be compared across residues within an aligned sequence position, the structural conformation and local 3D environment of each position can also be compared. Furthermore, structure

data provides knowledge of residue neighbors in 3D that may be distant sequentially. However, appropriately compensating for structure overrepresentation while being able to take advantage of the additional information provide by alternative structure conformations is a major obstacle. Overcoming the problematic structural overrepresentation of some protein sequences within an alignment while still incorporating *all* available structure data is a challenged specifically addressed by CCORPS.

4.4 Method overview

As input, CCORPS takes an aligned set of protein structures and a corresponding set of annotation labels (one label per structure). Then, CCORPS compares the structural and chemical similarity of subsets of the aligned positions across the alignment. Structural and chemical features common to structures having the same annotation label, while distinct from structures with differing annotation label, are detected as specificity-determining substructure instances. Finally, CCORPS predicts the annotation labels for structures with unknown annotation label and ranks the specificity-determining power of the alignment positions analyzed. For each specificity determining structural feature that is identified, as shown in Section 4.5, the positions responsible for the specificity are tallied. The alignment positions are finally ranked by their frequency of appearance in specificity-determining structural features as detailed in Section 4.8.

4.4.1 Computing residue position subset clusterings

In order to identify locally similar features among sub-groups of protein structures, all k-sized combinations of the r residue positions (i.e., $\binom{r}{k}$ combinations) are generated. For example, given $r = 20$ and $k = 3$, all $\binom{20}{3}$ 3-position subsets (1140 subsets) are generated.

Then, each of these position subsets are examined one-by-one. Continuing the example, given the position subset $(7, 13, 14)$, the protein structures are compared by examining the pairwise similarity of *only* positions 7, 13, and 14 in isolation (i.e., disregarding the other 17 positions).

The steps outlined in Sections 3.7, 3.8, and 3.9 are repeated for each possible 3-position subset in order to compare all possible local structural features across all proteins. Structural variation in most subsets is not expected to be informative, either because no significant variation is present, or because spurious patterns can occur due to chance. However, functionally relevant structural variation can be detected with many different subsets and therefore distinguished from random patterns, as will be shown below.

4.5 Selecting Highly Predictive Clusters (HPCs)

A cluster that is dominated by one annotation label can be used to predict the label for other structures in that cluster whose annotation is unknown. We therefore call such clusters Highly Predictive Clusters (HPCs). Identification of HPCs is performed by selecting a minimum threshold for the label purity of clusters, and then selecting all clusters with equal or greater label purity than this minimum as HPCs; we used the strictest purity threshold possible (1.0 or 100% purity) in this work (see Fig. 4.2). Purity is calculated for a multiset of labels, L, as purity$= I_L(mode(L))/|L|$ where I_L is the multiplicity function of a label within the multiset L and $mode(L)$ is the most frequent label within L.

An important consideration when evaluating the purity of clusters for HPC selection is the structural redundancy of the dataset. Overrepresented protein sequences will tend to form clusters due to being alternative structures of the same or highly similar protein sequences. A cluster of proteins made up of only a single or few highly similar sequences

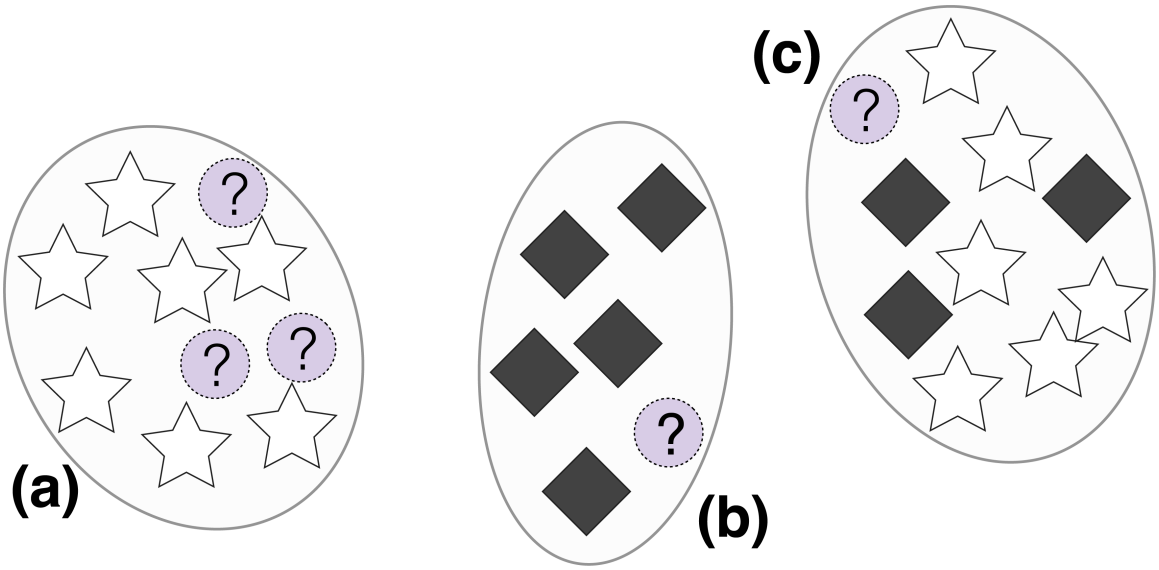


Figure 4.2 : **Illustration of cluster evaluation procedure.** The star and diamond symbols represent structures with known labels and the question marks represent structures with an unknown label. Clusters **(a)** and **(b)** will both be selected as HPCs for their respective labels (star and diamond, respectively) because they are each pure in a single label (unknown labels are disregarded). Cluster **(c)** will not be selected as an HPC because it has low purity.

is likely to exhibit label consensus by chance alone.

In order to eliminate label bias in clusters due to overrepresentation, a sequence non-redundant version of cluster purity was implemented. Protein sequences were clustered for all structures at the 100% sequence identity level for determining non-redundant groups in this case. To adjust purity for structural overrepresentation, NR-purity is calculated as NR-purity$= I_{L_{nr}}(mode(L_{nr}))$ where L_{nr} is one label from each NR $_{100}$ sequence identity cluster that exists within L.

Additionally, the purity of a set of labels must address the presence of proteins with unknown label in a cluster, which will occur because of the semi-supervised nature of CCORPS. In this work, proteins with unknown labels have no effect on the purity calculated for a set of labels. Therefore, we refer to the customized purity calculation used here as

semi-supervised, non-redundant purity or SS-NR-purity.

Finally, purity alone does not account for the distinctness of the proteins in the cluster relative to the remainder of the dataset. For example, an HPC for label l_1 that partially overlaps a second HPC for label l_2 is less likely to be informative than a l_1 cluster greatly separated from the remainder of the dataset. The "degree of separation" or "distinctness" of a cluster was quantified by calculating the cluster silhouette score [76]. The mean silhouette score for a cluster was then used as a further selection criteria for identifying HPCs by removing potential HPCs with negative average silhouette scores (malformed clusters).

4.5.1 Tallying votes for label predictions

Each time a protein falls within an HPC, that protein receives a single vote in favor of the majority label associated with the HPC. Because a protein can be a member of at most one cluster per k-position subset, the maximum number of votes any protein can receive is equal to the number of possible k-position subsets. For any given k-position subset, it is possible that all clusters are HPCs or that no clusters are HPCs depending on how the labels are distributed among the clusters. It is also possible that a protein may never fall within any HPC and therefore would receive zero votes for any label; such proteins are excluded from further analysis after the voting step. In the experiments described in Ch. 5 and 6 this case rarely occurred. After tallying the label votes across all k-position subsets, the label predicted for a given structure is then determined by selecting a decision boundary as described next.

4.6 Learning a decision boundary

Given a set of label votes that have been determined for an unlabeled structure, the threshold(s) used to decide which of the two or more label classes to assign to the structure requires the definition of a decision boundary procedure. For example, given a set of annotation labels containing the label classes $\{\texttt{true}, \texttt{false}\}$, a simple decision rule may be that given a structure with > 1 `true` vote, predict the `true` label for that structure. However, determining a single threshold for deciding the number of label votes required to classify a structure into one of several classes is difficult to generalize. Two approaches to selecting a label prediction given the label votes for unlabeled structures are described below. The two approaches were used in the applications of CCORPS in Ch. 5 and 6, respectively. While the two approaches below were both proven effective for the applications demonstrated in this thesis, a variety of supervised learning approaches could instead be applied to determine the decision boundary (e.g., *k*-nearest neighbors, decision trees).

Majority vote

One simple alternative to selecting a static threshold for the number of votes required to classify into a particular label class is to simple select the class that has the most votes, regardless of the number of votes cast for other classes, individually or combined. (i.e., the predicted label is the mode of the label votes). For example, given a structure with the label votes {85 `true`, 1032 `false`}, a majority vote procedure would predict the `false` label for the structure. In Ch. 5, this straightforward approach to predicting labels is illustrated to be successful for selecting among k EC class labels.

*Support Vector Machine-based (*SVM*) decision boundary*

However, the majority vote approach described above decays in predictive performance when the label class distribution among the input substructures is highly non-uniform. For example, if most proteins in the dataset are of label `false` and only a small fraction have label `true`, then many HPC`false` clusters will be identified by chance alone because the `false` label dominates the dataset. Simple majority voting will result in large numbers of false negative predictions due to the abundancy of the `false` label.

Because CCORPS is a supervised approach, the labels for the training structures are known and can be used to empirically estimate a vote count decision boundary. For example, given structure X with known label, the number of times that X appeared in a HPC`false` and HPC`true` across all k-position subsets can be calculated using the same approach as for unlabeled structures. The structure X is then represented by an $|l|$-dimensional vote vector, where each of the l dimensions corresponds to the number of votes X received for label l_i, $1 \leq i \leq l$. Application of this procedure to all labeled structures in the dataset provides an empirical basis for calculating a decision boundary in the vote space given the vote distribution for labeled structures. For example, the blue and red points shown in the scatter plot of Fig. 4.3 denote the vote vectors for training set substructures with known `true` and `false` labels, respectively.

Given the vote vectors calculated for all labeled training set substructures in the dataset, it is then possible to then train any number of classifiers in order to determine a decision boundary. To compute a decision boundary in the vote space for classifying unlabeled proteins, SVMs were selected. First, an SVM (linear kernel) is trained using the vote vectors of labeled training set substructures. For example, the decision boundary determined by training an SVM on vote vectors is shown in Fig. 4.3 as the bold, black line. Next, for an unlabeled substructure with a given vote vector, the label for the substructure can be pre-

dicted by determining which side of the SVM decision hyperplane the unlabeled structure falls within. As illustrated in Fig. 4.3, test vote vectors falling within the blue region will be predicted as having the `true` label and those falling within the red region, the `false` label.

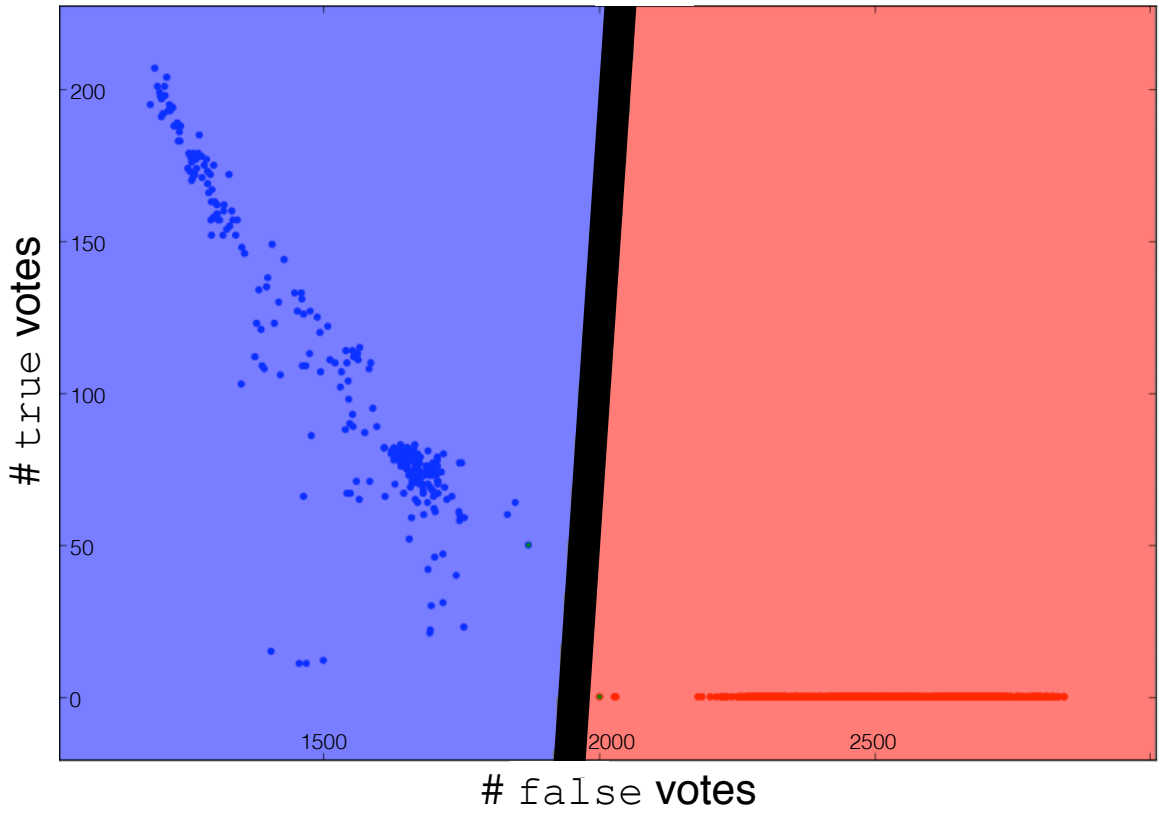


Figure 4.3 : **Decision boundary for label vote vectors computed by SVM.** In the above scatter plot, each point corresponds to the number of `true`/`false` votes accumulated by each substructure across all clusterings. Combining the above label vote vectors with the known labels for substructures to train an SVM (using linear kernel) results in the decision boundary shown as the bold black line. The red and blue regions (right and left sides of the boundary, respectively) denote the values for which the predicted label will be `false` and `true`, respectively. Blue points indicate substructures known to have the `true` label while red points denote the `false` label. In the case of Roscovitine above, wide separation between the two classes exists.

4.7 Cross-fold validation

To assess the utility of HPCs for identifying substructure positions indicative of functional specialization, a cross-fold validation can be performed as follows. First, the input substructures are divided into 95% sequence identity groups (NR_{95}-clusters), and each of these NR_{95}-clusters will be one fold. This procedure ensures no protein in a test set shares >95% sequence identity with any protein in the training set for any given fold of the evaluation. Because of the non-uniform distribution of structures across the NR_{95}-clusters, the number of structures in the test set varies with each fold. In each fold, structures that were part of the test set are marked with label `unknown`, and are disregarded when calculating the SS-NR-purity of clusters during the HPC selection step, just as the structures with truly `unknown` label.

Finally, standard k-fold cross validation was performed with each of the NR_{95}-clusters each being one fold (i.e., $k = |NR_{95}\text{-clusters}|$). Given the NR_{95}-clusters-based fold partitioning above, the training set is used to identify HPCs and then tally label votes for each structure in the test set of a fold. For a given test structure, the predicted label is compared to the actual label and the result is tallied in a $m \times m$ confusion matrix, where m is the number of unique labels.

4.8 Ranking of specificity determining positions

Each identified HPC represents a k-position subset of the alignment for which some subset of the proteins within the family are structurally distinct from the remainder of the family and all share a common annotation label; see Fig. 5.2 for examples of HPCs. Each alignment position is ranked by the number of times that position belonged to a k-position subset containing one or more HPCs; note that $\binom{r}{k}$ k-position subsets of r residue alignment

positions exist. Tallying the number of HPC appearances for each position across all $\binom{r}{k}$ clusterings results in a per-position HPC tally. See Section 5.7 for additional details on the ranking of SDPs for Pfam binding site positions.

4.9 Conclusion

The CCORPS method introduced here is completely automated and can be applied to a variety of binding site analysis problems. The generality of the approach means that it can be used for any kind of annotation label. The approach is also agnostic with respect to the method used for alignment in similar fashion to The FASST Framework. In the following chapters, the application of CCORPS to the prediction of different sets of annotation labels, such as enzymatic function specialization and kinase inhibitor binding affinity, will be demonstrated.

To validate the predictive ability of the structural features identified by CCORPS an extensive dataset of 48 families was automatically constructed using the Pfam database [77] as a source of well-curated protein alignments. These benchmarking results are discussed in Ch. 5. For the benchmarking experiments in Ch. 5, the annotation labels analyzed are per-structure Enzyme Commission (EC) number classifications. Cross-validation is performed in order to evaluate the predictive power of CCORPS and the utility of the distinguishing structural features identified.

Beyond the benchmark Pfam+EC dataset presented here, CCORPS provides a general framework for automatically learning structural features that distinguish proteins having different annotation labels. A major advantage of the work presented is the generality of CCORPS to detect structurally distinguishing features for a wide variety of applications beyond enzymatic function specialization. No assumptions regarding the nature of the

annotation labels nor of the alignment type are made at any point by CCORPS. That is, a phylogenetic relationship is neither assumed nor required for input protein alignments. This allows the incorporation of purely structure-based alignments, such as those available in databases like HOMSTRAD [78] or even local structure alignments such as those identified by motif/template search algorithms (e.g., SOIPPA, [62] and LabelHash, [2]). Other sources of annotation labels, including Gene Ontology (GO, [63]) terms, binding affinity for a given molecule and ligation state can be incorporated as-is with CCORPS without modification to the method.

Chapter 5

Predicting Enzymatic Classifications for Protein Domains

5.1 Motivation

Large-scale analysis of the structural variation of protein enzymatic sites has become increasingly important in identifying the role of specific structural features responsible for functional diversification [26]. However, the inherent difficult of combining all of the available structure data for a given protein sequence in order to consider all of the possible binding site conformations present among structures with non-cognate ligands, transition state analogs and also the apo, unbound state has proven difficult [33, 68].

Recent work [3, 79] has demonstrated that enzyme superfamilies tend to share a common catalytic core that is then embellished by the insertion of additional secondary structure elements and specificity loop substitutions that give rise to functional diversification. Comparing the binding site specific structural variation present among these families of homologous proteins has been shown to identify distant structural relationships that are

sequence-based approaches fail to recognize [16].

In this chapter, the ability of CCORPS to successfully identify structural features that are predictive of enzymatic specialization in families of homologous proteins is demonstrated on a large protein family benchmark dataset for multiple levels of enzyme functional classification specificity.

After introducing the Pfam binding site dataset automatically constructed to benchmark CCORPS, we will introduce three distinct results of the method. First, we will discuss the accuracy of CCORPS in predicting EC classifications in a large-scale, cross-fold validation experiment. Second, we will demonstrate that HPCs are capable of distinguishing structures with differing EC classifications and that multiple HPCs can exist for a given EC class. We conclude this section with a discussion of the identification of specificity determining positions from HPCs.

The overall classification accuracy of CCORPS (Table 5.1) when applied to the Pfam+EC dataset demonstrates the ability of CCORPS to identify structural features that distinguish functionally different protein homologs.

5.2 Related work

A variety of approaches for identifying SDPs have been successfully demonstrated before. Methods such as GroupSim [80], Xdet [81], MCdet [81] and ET [82] identify column positions within multiple sequence alignments (MSAs) that distinguish functional sub-families. See [80] for a review of sequence-based approaches for SDP identification.

The recent FLORA [33] method uses a structure-based alignment approach called CATHEDRAL [48] to identify positions within domains that are structurally conserved among proteins sharing a common function. However, to the best of our knowledge, no method is

currently able to incorporate and examine all structure data available for a given homologous protein domain when identifying SDPs.

5.3 Dataset

All 12,273 protein families from the Pfam 25.0 release (April 2011) were considered for inclusion in the dataset. Only protein families that met the following criteria were selected for inclusion: ≥ 200 domain structures; ≥ 10 unique sequences, ≥ 2 distinct EC classes; for the subset of sequences with known EC class, ≥ 2 sequences from each of ≥ 2 EC classes, all having $\leq 50\%$ sequence identity. Our dataset consists of the 48 protein families that meet or exceed these criteria. The criteria were chosen so that there would be enough structural and sequence diversity to make the prediction of EC classifications sufficiently challenging.

5.3.1 Automated binding site definition

All alignments used in this work were derived from Pfam MSAs [77]. A Pfam MSA provides an alignment of homologous protein domains. For each aligned domain in an MSA, the UniProt [83] ID is retrieved from the Pfam alignment and all PDB [17] structures corresponding to the given UniProt ID are mapped to the domain sequence. The Pfam MSA alignment column positions define the mapping of residue positions across all structures for a protein family. Examples of the binding site definitions automatically generated for our dataset are shown in Fig. 5.1.

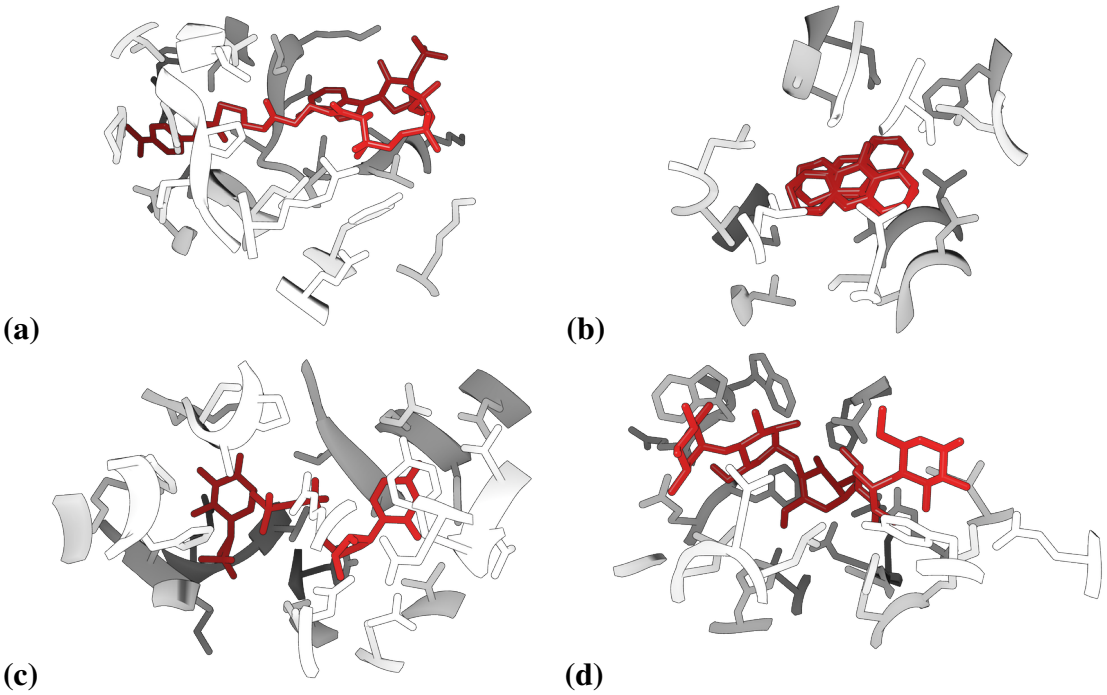


Figure 5.1 : **Binding site positions for some of the protein families analyzed by CCORPS.** A representative structure is shown for each family. The automatically selected binding site residues are shown in white, while the ligand is shown in red. **(a)** ECH, **(b)** COesterase, **(c)** Epimerase, and **(d)** Alpha-amylase.

5.3.2 Selecting binding site positions

For each aligned structure that contains one or more non-protein molecules (distinguished by HETATM records) with ≥ 30 atoms, the largest available molecule was identified and assumed to be a ligand. For each ligated structure, all residues having at least one atom within 5Å of one or more ligand atoms were selected as potential binding site residues. These binding site residues were then mapped to columns within the Pfam MSA. A count is kept for the number of times each MSA column was mapped to a residue in a ligated structure. After tabulating MSA column mapping counts across all ligated structures, only MSA columns that were mapped to binding site residues in ≥ 5 instances were retained.

5.3.3 Identifying a dense sub-alignment

Next, it is necessary to remove gaps from the input alignment so that all pairwise comparisons of binding site positions, as outlined in Section 3.7, are consistent. When gaps appear in the aligned binding site column positions, either the entire column position must be eliminated from further analysis or all protein structures having a gap at the alignment position must be eliminated. This "densification" procedure of removing either a gapped row (protein structure) or gapped column (alignment site position) is repeated until only a fully "dense" (non-gapped) sub-matrix remains. In the resulting dense sub-matrix, all remaining protein structures have a residue at all remaining alignment positions.

Finding the largest dense sub-matrix in the alignment as outlined above is equivalent to finding a maximal edge biclique in a bipartite graph; a biclique is a complete bipartite graph, where all possible edges exist between two bipartite sets of vertices. Given a graph $G = (V_1 + V_2, E)$, alignment positions are vertices in V_1 and protein structures are vertices in V_2. Each non-gapped position for a protein $v_i \in V_1$ at a position $v_j \in V_2$ in the alignment is the edge $E = v_1 v_2$. Identifying the maximal biclique in such a bipartite graph has been shown to be NP-complete [84].

The heuristic densification approach implemented to identify dense sub-matrices of alignments is as follows. (1) Given n structures aligned (with gaps) at m positions, convert the alignment to an $n \times m$ binary matrix M such that $M[i][j] = 0$ if structure i was gapped in the alignment at position j and $M[i][j] = 1$ otherwise. (2) Consider each row $M[i]$ to be a binary vector representing structure i. (3) Compute the complete-linkage hierarchical clustering [85] of the binary vectors using the Hamming distance metric [86]. (4) Each node of the resulting hierarchical clustering represents one potential sub-matrix. Calculate the dense size of the sub-matrix by removing all rows or columns containing one or more zeros from the sub-matrix and taking the sum of the remaining values. (5) Select the sub-matrix

with maximal dense size. Note that the maximal dense sub-matrix selected in step (5) is not guaranteed to be the optimal sub-matrix because every possible sub-matrix of the original matrix does not exist as a node in the hierarchical clustering. The rows (protein structures) and columns (alignment positions) for the selected dense sub-matrix are used to prune the raw alignment positions and structures in order to provide a fully dense "sub-alignment" as input to CCORPS.

Techniques used for finding dense sub-matrices within real-valued gene expression data such as "biclustering" are potential alternatives to the heuristic approach used here (see [87] for a review of biclustering approaches). However, the binding site position subset of an alignment is often quite dense, making the sparseness assumptions of gene expression biclustering methods unessential for the current Pfam alignment dataset.

5.3.4 Generating EC class annotation labels

For each protein structure in a family, several different annotation labels are generated based upon the 4 tiers of EC classification. For example, a given structure with an EC classification of the form `A.B.C.D` can be labeled for each tier of the EC as `A.B.C.D`, `A.B.C.*`, `A.B.*.*`, or `A.*.*.*`. The 4-tiered label (`A.B.C.D`) provides a more precise functional label than the 1-tiered version (`A.*.*.*`). The objective is to predict all 4 EC labels for structures with unknown EC classification.

5.4 Problem definition

Given the protein family dataset described in Sec. 5.3, the predictive ability of CCORPS for the identification of structural features specific to functional divergence among homologous enzyme active sites was assessed. The aligned substructures generated for each Pfam fam-

ily were independently provided as input to CCORPS along with the EC annotation labels for each tier.

For each Pfam family, and each EC annotation label tier, individually, the aligned substructures and annotation labels were provided as input to CCORPS and the cross-fold validation procedure described in Sec. 4.7 was performed. That is, 42 (# families) $\times$ 4 (# EC labelings) independent CCORPS cross-fold validations were performed.

Table 5.1 : **Accuracy of predicted EC classifications for Pfam protein families in cross-fold validation.** Predictions are made at all 4 tiers of the EC hierarchy.

	Family statistics			**EC prediction accuracy (%)**			
Pfam ID	#EC	#Struct.	Size*	1-tier	2-tier	3-tier	4-tier
4HBT	5	37	28	100	95	95	10
AAA	4	57	14	30	26	26	26
ADH_N	10	62	9	100	99	99	68
Aldedh	15	66	58	100	90	90	71
Alpha-amylase	16	220	26	89	89	89	71
Amino_oxidase	7	51	33	100	12	12	6
Aminotran_1_2	15	148	32	96	95	95	39
Asp	11	54	75	100	100	100	14
COesterase	6	152	62	100	100	100	78
Cu-oxidase	7	95	11	100	88	49	49
DHFR_1	5	155	81	94	94	94	92
ECH	10	38	43	81	79	79	36
Epimerase	9	62	71	95	95	95	88
Ferritin	4	79	15	90	90	2	2

Table 5.1 : **Accuracy of predicted EC classifications for Pfam protein families in cross-fold validation.** Predictions are made at all 4 tiers of the EC hierarchy.

	Family statistics			**EC prediction accuracy (%)**			
Pfam ID	#EC	#Struct.	Size*	1-tier	2-tier	3-tier	4-tier
GST_C	5	209	5	91	91	91	91
GST_N	4	134	18	98	98	98	98
Glyco_hydro_18	3	112	16	100	100	98	98
Gp_dh_C	4	47	11	100	100	100	88
Hexapep	13	76	13	90	75	75	48
Lactamase_B	9	119	4	98	44	40	40
Ldh_1_C	4	112	3	100	100	100	58
Ldh_1_N	4	64	43	100	100	100	86
Lys	8	548	22	100	100	100	96
NUDIX	9	33	41	90	88	85	24
PALP	18	113	9	72	52	52	44
PDEase_I	5	98	28	100	100	100	82
Peptidase_C1	11	100	54	99	99	99	16
Peptidase_C14	6	19	44	100	100	100	45
Peptidase_M10	12	131	64	100	100	99	30
Peptidase_S9	7	100	12	100	99	97	94
Pkinase	11	145	69	100	100	57	52
Pkinase_Tyr	5	142	66	100	100	66	52
Proteasome	3	20	64	100	100	46	46
Proteasome_A_N	3	21	7	100	100	47	47

Table 5.1 : **Accuracy of predicted EC classifications for Pfam protein families in cross-fold validation.** Predictions are made at all 4 tiers of the EC hierarchy.

	Family statistics			**EC prediction accuracy (%)**			
Pfam ID	#EC	#Struct.	Size*	1-tier	2-tier	3-tier	4-tier
Pyr_redox	21	145	14	100	65	65	57
Pyr_redox_2	11	49	76	100	59	59	10
Pyr_redox_dim	12	94	7	100	56	56	38
RVP	9	418	35	99	99	97	93
Ribonuclease	6	158	14	100	100	91	52
Rieske	5	63	33	100	96	96	74
TPP_enzyme_C	11	59	21	83	83	56	48
TPP_enzyme_N	14	78	17	76	81	52	42
Thioredoxin	12	171	3	44	0	0	0
Thymidylat_synt	5	106	45	92	92	92	86
adh_short	21	137	65	100	72	72	35
efhand	14	308	17	15	15	1	0
p450	7	67	43	100	100	18	16
peroxidase	3	45	63	100	100	100	100
Mean				92	84	74	53
Standard Deviation				18	25	30	30

5.5 Prediction performance

The protein family dataset that we have constructed in the manner described above covers a wide range of families with very different levels of functional diversification and binding site sizes as shown in Table 5.1. The mean number of unique EC classes across families in the dataset was 8.3, with some families having as few as 3 different EC classes and as many as 21. An even wider variance is seen for the number of structures available per family, ranging from as few as 11 to as many as 548 with a mean of 108 for the dataset. Finally, the number of binding site positions examined ranged from the minimum of 3 to as many as 81 (mean of 33), covering a large range in binding site sizes.

The performance of CCORPS was evaluated by applying the cross-validation procedure outlined in Section 4.7 to each of the protein families in the Pfam alignment dataset. For each protein family, the ability of CCORPS to predict enzymatic function annotation labels in the form of EC class numbers was quantified. The prediction accuracy of CCORPS for predicting EC classification at each of the 4 tiers of EC specificity is show in Table 5.1.

Due to the hierarchical nature of the EC classification system, the number of unique 4-tier EC classes (most specific annotation labels) for a family is necessarily greater than or equal to the number of unique 1-tier EC classes (least specific annotation labels). As can be noted by examining the dataset mean prediction accuracy from 1-tier to 4-tier, accuracy decreases with increasing EC classification annotation label specificity, as should be expected. The prediction accuracy at the least specific 1-tier EC classification was $92 \pm 18\%$, while the accuracy dropped to $53 \pm 30\%$ for the most specific 4-tier. With these numbers one needs to consider the very general automated procedure used to specify the input (e.g., the way binding site residues were chosen).

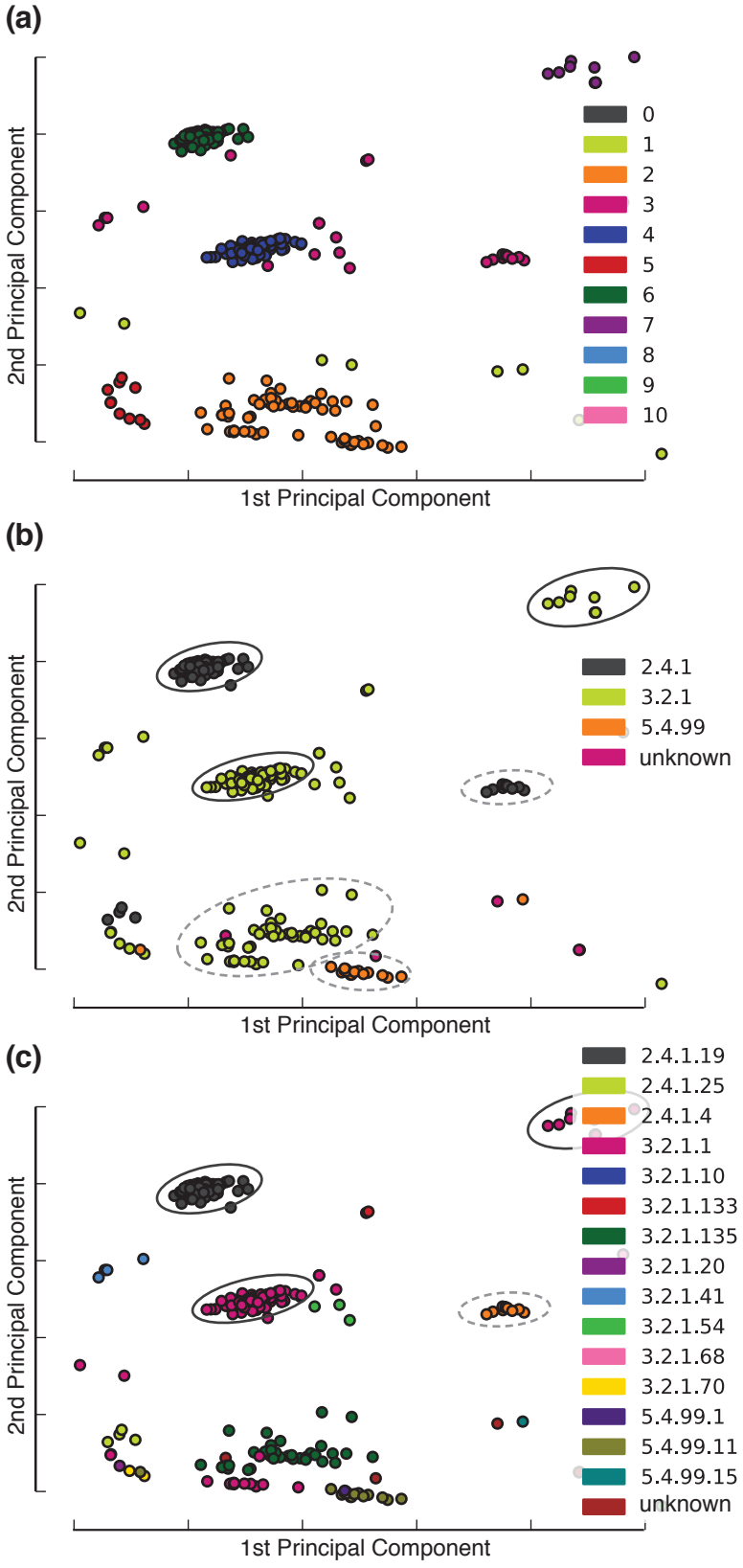
A major challenge when attempting to predict the 4-tier (most specific) EC annotation

labels derives from the non-uniformity in structure coverage across the EC labels within a protein family. What could be considered "outlier" EC classes with only a single corresponding protein sequence within a protein family, were common throughout the dataset. Given the stringent cross validation procedure used in this work, it is actually impossible to correctly predict the annotation label for single-protein EC classes. This is due to the fact that when the structures for the single-protein EC class fall within the test set during one fold of the cross validation, no structures will exist in the training set that share the same EC class label by definition of the NR-clusters. We chose to be conservative and not correct for this self-penalizing aspect of our performance benchmarking. The reason for this is that it reflects the realistic case of predicting enzymatic function for homologous proteins with unknown function that may be novel relative to the training dataset.

5.6 Highly predictive clusters

The basis for the predictive ability of CCORPS is the detection of HPCs as outlined in Section 4.5. The set of clusters identified by CCORPS for one of the 2600 3-position subsets for the α-amylase family is shown in Fig. 5.2. Note that $\binom{26}{3} = 2600$, where 26 is the number of binding site positions available for the α-amylase protein family as listed in Table 5.1.

In the subplots of Fig. 5.2, the points shown each represent a feature vector (as calculated in Section 3.7), where each feature vector corresponds to a single protein substructure. A tightly grouped cluster of feature vectors in the subplots of Fig. 5.2 reflects a set of substructures sharing a high degree of structural and chemical similarity. Fig. 5.2**(a)** shows the cluster membership automatically identified by CCORPS. Fig. 5.2**(b)** and **(c)** show the EC annotation labels that map to each feature vector at the 3-tier and 4-tier levels, respectively. In the last two plots the points are colored by EC tier level rather than by cluster.


(a)
2nd Principal Component
1st Principal Component
0
1
2
3
4
5
6
7
8
9
10
(b)
2nd Principal Component
1st Principal Component
2.4.1
3.2.1
5.4.99
unknown
(c)
2nd Principal Component
1st Principal Component
2.4.1.19
2.4.1.25
2.4.1.4
3.2.1.1
3.2.1.10
3.2.1.133
3.2.1.135
3.2.1.20
3.2.1.41
3.2.1.54
3.2.1.68
3.2.1.70
5.4.99.1
5.4.99.11
5.4.99.15
unknown

As can be seen in Fig. 5.2**(b)** and **(c)**, several HPCs for different labels can exist simultaneously in a single clustering. Also, as shown in Fig. 5.2**(b)** for the 3-tier EC label `3.2.1`, multiple distinct HPCs for a single label can be identified. In other words, within EC label `3.2.1` several structural sub-groups can be detected. The existence of distinct HPCs for a single label indicates that multiple structurally and chemically distinct sub-groups can exist within a common annotation label. It is possible to identify such instances because CCORPS makes no assumptions about the structural homogeneity of sub-families having the same enzymatic function.

5.7 Identifying specificity determining positions

Because the residues which distinguish an EC class from the remainder of the family are not necessarily the same for all EC classes within a protein family, SDPs are ranked separately for each label.

Specificity determining binding site positions that distinguish a sub-family from other homologous proteins with differing function are determined by constructing a relative ranking of all binding site positions using the procedure outlined in Section 4.8. As shown

---

Figure 5.2 *(preceding page)*: **Substructure clustering for one 3-position subset of the α-amylase binding site alignment.** In each scatter plot above, the dimensionality-reduced feature vectors computed by CCORPS are shown. Each point shown is one feature vector and each feature vector represents one protein substructure. Tightly grouped points correspond to binding site substructures with high structural and chemical similarity. Plots **(a)**, **(b)** and **(c)** above all show the same clustering with different sets of annotation labels applied (labels are denoted by color): **(a)** cluster ID labeling; **(b)** 3-tier EC labeling; **(c)** 4-tier EC labeling. Solid ellipses indicated clusters identified automatically as HPCs. Dashed ellipses indicate subsets of non-HPC clusters that would have been considered HPCs if the clustering step had distinguished each as a separate cluster.

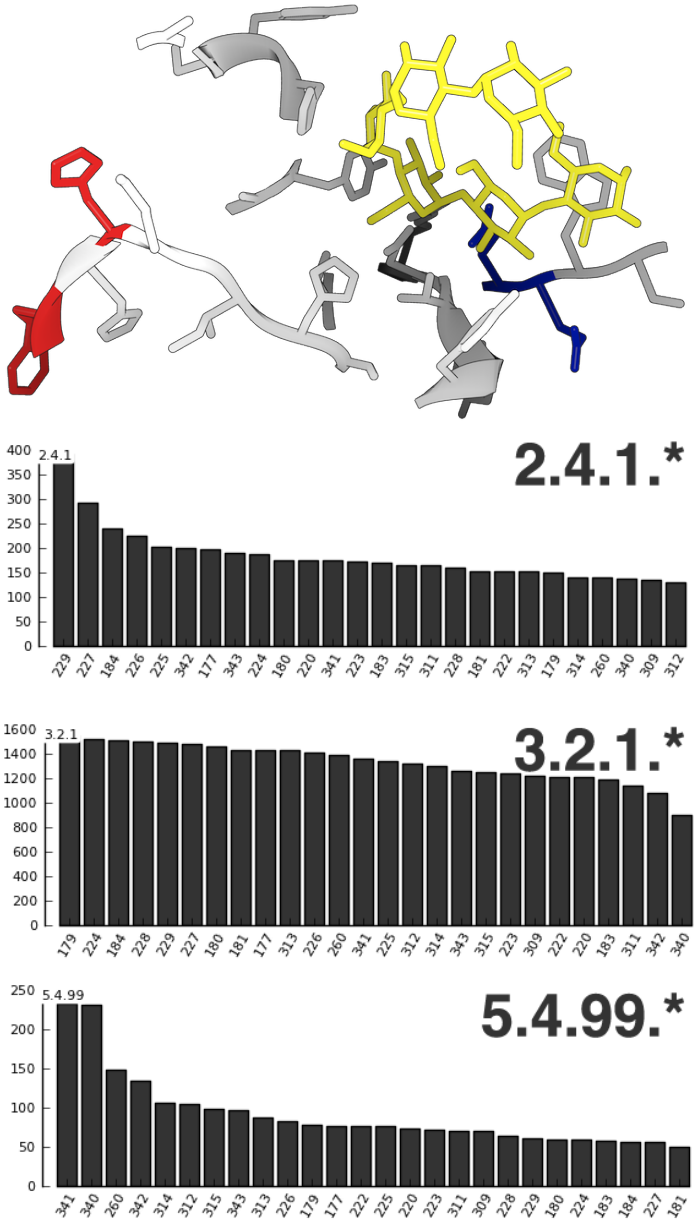


Figure 5.3 : **Predicted binding site specificity determining positions for the α-Amylase family.** The α-Amylase binding site positions are shown in stick above with bound ligand in yellow, high-ranking SDPs for EC `2.4.1.*` (229 and 227) are colored red, high-ranking SDPs for EC `5.4.99.*` (341 and 340) are colored blue. For the 3 bar charts, the residue number for each alignment position considered is shown along the x-axis (residue numbering is according to PDB:3EDF); the number of instances where the position corresponded to a clustering containing one or more HPCs is shown on the y-axis. The positions are sorted along the x-axis in order of specificity determining power. Note that no prominent alignment positions were identified for EC `3.2.1.*` in this case.

in Fig. 5.3 for the tier-3 EC annotation labels and the α-amylase protein family, a separate ranking is constructed for each EC class. The position rankings for EC `2.4.1.*` and `5.4.99.*` reveal that positions 229 and 227 for `2.4.1.*` and positions 340 and 341 for `5.4.99.*` appear much more frequently in HPCs for their respective labels than the remaining binding site positions. In contrast, all positions occur with similar frequency within HPCs for label `3.2.1.*`, which seems to indicate that the binding sites of proteins in EC `3.2.1.*` as a whole differ quite significantly from the binding sites of all other proteins in the family.

5.8 Conclusion

As demonstrated in this chapter, CCORPS is able to identify structural features that are predictive of enzymatic function specialization within families of homologous proteins. Additionally, the predictive structural features can be used to evaluate the contribution of individual binding site residue positions towards determining functional specificity. Many of the component steps of CCORPS can be further refined in order to improve the predictive performance of CCORPS. For example, the initial binding site position selection methodology (Section 5.3.1) has a drastic impact on the performance of CCORPS, because if a true SDP is not included among the positions provided to CCORPS, it will of course not be identified. While determining a binding site region for a single ligated structure is very well-defined by simply selecting all residues within a distance cutoff of the ligand, defining a true "consensus binding site" across a large, heterogeneous set of structures, in general, is not well-defined due to the lack of a complete one-to-one mapping among all secondary structure components for the set of structures.

While the SDP ranking presented here can identify prominent positions in the case of

some EC classes, as illustrated with the α-amylases in Fig. 5.3, determining a statistical cutoff for distinguishing SDP s from non-SDP s based upon the rankings is challenging. This will require the development of a statistical model that properly incorporates the prior probability of HPCs based upon the distribution of annotation labels within a protein family as well as the likelihood of HPCs occurring by chance alone.

CCORPS provides a local structure-based approach to identifying potential SDP s and predictive structural features for enzymatic function specialization. The redundancy-aware methodology implemented by CCORPS allows for *all* available structure data to be incorporated without prior filtering or redundancy removal and systematically accounts for structural overrepresentation within each method step. The large protein family dataset analyzed here demonstrates the success of the method across a wide variety of different binding site and family sizes, with many families having a very large number of unique EC class annotation labels to predict. As will be further demonstrated in the following chapter, CCORPS can be applied to make even more fine-grained, molecule specific binding predictions for families exhibiting large amounts of structural conservation within the binding site region analyzed.

Chapter 6

Predicting Binding Affinity for the Human Kinome

6.1 Motivation

The protein kinases constitute the largest enzyme family encoded by the human genome, with currently 518 known instances, making up 1.7% of all human genes [9, 4]. Because these protein kinases are intimately involved in cellular communication and regulation networks, the loss of normal kinase regulation has been implicated in a wide variety of pathological conditions. The number of disease states found to be associated with kinase dysregulation has motivated the development of kinase-specific inhibitor compounds. Because of the number of kinase-associated conditions, the protein kinases have come to constitute 20-30% of the drug development programs at many companies [9].

Due to the large number of existing protein kinase domains and the high degree of (ATP) binding site similarity among them, designing highly selective inhibitors has proven difficult. For example, type I kinase inhibitors that only target the ATP site have typically

been found to have low selectivity across the kinome [10]. To increase inhibitor selectivity, type II inhibitors bind both the ATP site and the immediately adjacent allosteric site. By also binding to the allosteric site, type II inhibitors are able to make additional highly specific interactions, thereby allowing them to be more selective [10]. However, identifying highly specific structural features that can be uniquely targeted by inhibitors requires the comparison of local structural similarity across kinase binding sites.

Many of the effective inhibitor selectivity strategies involve exploiting the differences in the size of the ATP binding site and targeting residue variability at a few key positions [10, 11]. These structure-based comparison approaches have proven more useful than sequence-only measures of overall kinase similarity in evaluating the potential selectivity profile of inhibitors [11]. For example, the size of the gatekeeper residue directly moderates the availability of the hydrophobic pocket. Inhibitors having larger functional groups that bind the hydrophobic pocket may be able to select for the roughly 20% of protein kinases that have a relatively small gatekeeper residue (e.g., Gly, Val, Ala or Thr), because kinases having a larger gatekeeper residue (e.g., Phe) do not have a large enough hydrophobic pocket to accommodate the inhibitor [11]. However, in order to select for an even more specific subset of the human kinome, it has proven necessary to take advantage of multiple structural features of the kinase binding site (both ATP and allosteric sites) simultaneously.

Because of the importance of identifying trends among multiple structural features for a variety of different kinases, a comparative analysis of structure-affinity relationships within the kinome is presented here. By incorporating all available structural data for each kinase binding site, such as the variety of binding conformations that exist for with multiple ligated compounds, CCORPS is demonstrated to be capable of predicting the binding ability of kinase inhibitors across the human kinome.

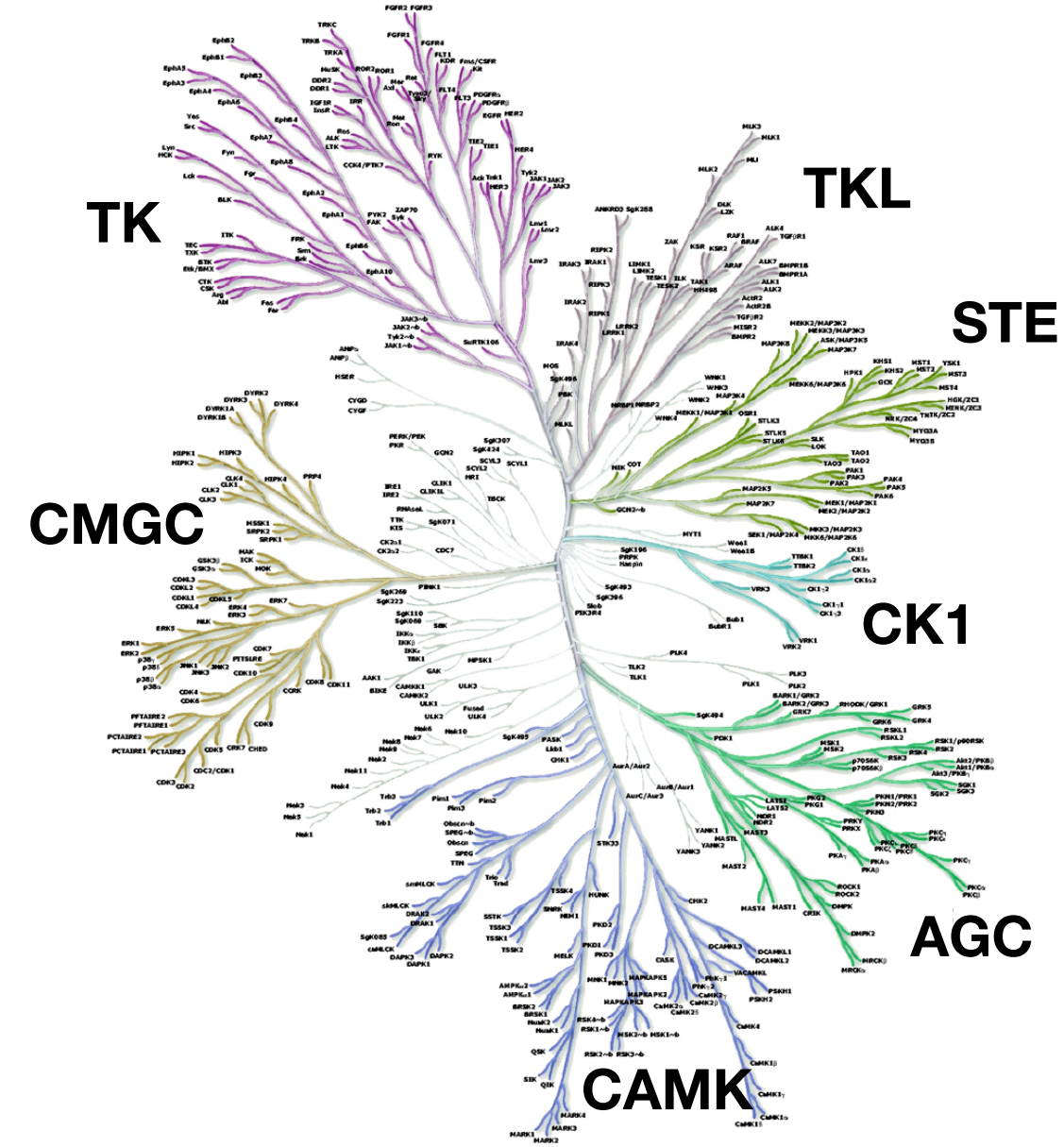


Figure 6.1 : **Phylogenetic tree of the human kinome.** The following seven major families make up the major branches of the kinome as shown above: containing PKA, PKG, PKC (AGC); Calcium/calmodulin-dependent protein kinase (CAMK); Casein kinase 1 (CK1); CMGC Containing CDK, MAPK, GSK3, CLK families; STE Homologs of yeast Sterile 7, Sterile 11, Sterile 20 kinases; TK Tyrosine kinase; TKL Tyrosine kinaselike [4]. The kinase dendrogram was adapted and is reproduced with permission from Science (http://www.sciencemag.org) and Cell Signaling Technology, Inc. (http://www.cellsignal.com)

6.2 Related work

The release of kinome-wide affinity screening datasets, such as that of Karaman et al. 2008 [6] and Fabian et al. 2005 [88], provides a novel opportunity to identify structural trends as they relate to binding affinity across kinases. Recent work [89, 90] has illustrated that local structural similarity exists among phylogenetically diverse groups of kinases and have highlighted the importance of large-scale, multiple-structure analysis of structure-affinity relationships among the kinases.

Recent work by Jackson et al. 2009 demonstrated a structural clustering approach to predicting kinase inhibitor binding affinities. Their geometric hashing approach to whole-site comparison of the ATP binding pocket was demonstrated to be effective at identifying possible instances of inhibitor cross-reactivity and further emphasized the importance of taking into account subtle conformational changes in the binding site.

Instead of whole-site comparison, recent work by Huang et al. 2010 [11] utilized a knowledge-based approach to constructing a minimal binding site "fingerprint" that captures only a pre-specified set of well-studied structurally selective features, such as the size and hydrogen-bonding ability of the gatekeeper residue.

However, even given the successes of existing approaches for predicting kinase affinity, several outstanding problems remain. The whole-site-based affinity prediction method of Jackson et al. 2009 relies upon the selection of a reference structure, known to bind a particular inhibitor, against which the similarity of all other sites are ranked. As noted in their prediction results for imatinib affinity prediction, the performance varied greatly depending on the particular structure selected as a reference. The approach used by CCORPS, however, provides an automated way to incorporate the similarity of an unannotated structure to all annotated structures without the need to manually select a reference. Because of the

variability in binding site conformation (e.g., DFG motif), poor reference structure selection can have drastic effects upon the identified structural similarity.

The approach implemented by CCORPS here allows not only the effective prediction of binding affinity across a variety of small molecule inhibitors, but as a by-product of the learning process, also identifies many instances of structurally similar features among phylogenetically diverse kinases, as will be demonstrated below.

6.3 Dataset

The kinome structural dataset was constructed from all structures annotated as belonging to PFAM:PKINASE and PFAM:PKINASE_TYR (all ePK domains, aPK s excluded) in release 25 of Pfam. After the binding site residue positions to analyze were selected (see Sec. 6.3.2) and proteins having one or more gaps at those positions were excluded, a total of 1958 structures remained. These 1958 structures correspond to 208 unique kinase proteins. The distribution of sequences and structures across the seven major kinome families is shown in the table below:

	AGC	CAMK	CK1	CMGC	Other	STE	TK	TKL	Unclassified
# Structures	171	231	20	500	114	55	445	58	364
# Sequences	19	34	6	33	18	17	47	9	75
# Annotated	6	13	2	16	5	11	35	6	43

As will be detailed in Sec. 6.3.3, 1281 of the 1958 structures (65.4%) were part of the kinome inhibitor affinity dataset of Karaman et al. 2008 and therefore have known annotation labels for each of the 38 inhibitors tested.

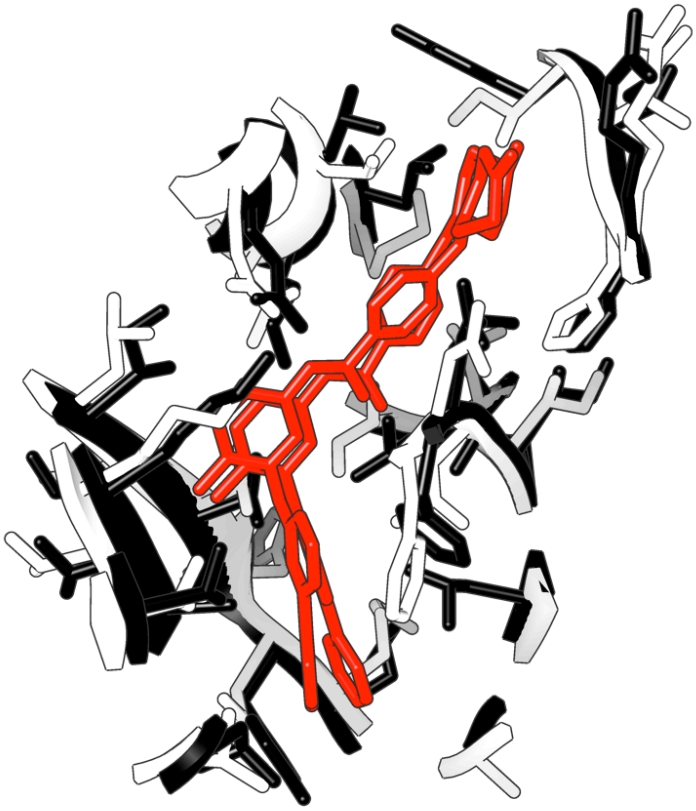


Figure 6.2 : **Structure-based binding site alignment via MATT.** In order to identify a mapping between residues in the TK and non-TK Pfam alignments, MATT [5] was used to compute a structural alignment of the kinase domains of p38 structure PDB:3HEC (white) and LCK structure PDB:2PL0 (black), both with bound imatinib inhibitor (red). The C_α RMSD of the above binding site alignment region (27 residue positions) was 1.169 Åand the RMSD of the imatinib inhibitors is 1.736 Å; the imatinib molecule coordinates were ignored during computation of the alignment.

6.3.1 Eukaryotic protein kinase alignment

In the case of the ePKs, both the non-TKs (PFAM:PKINASE) and TKs (PFAM:PKINASE_TYR) were combined into a single, comprehensive structural dataset in order to provide structural coverage for the full kinase family tree. However, determining an appropriate residue position correspondence *between* the TK and non-TK family alignments requires an additional alignment step, in order to relate columns from the TK alignment to columns of the

non-TK alignment. Several approaches for obtaining a consistent and high-quality alignment between the TKs and non-TKs were considered, such as profile-profile alignment and structure-based alignment.

To provide a structure-based solution to determining a high-quality residue position correspondence between the TKs and non-TKs, MATT [5] (version 1.00) was selected due to its ability to focus the alignment on regions of structural similarity (e.g., the ATP binding site) while disregarding regions with low structural similarity (e.g., C-terminal region). The kinase domain of non-TKs and TKs was then aligned using MATT by structural superposition of a pair of representative structures (PDB:3HEC and PDB:2PL0, respectively), that had both been co-crystallized with the same ATP binding site inhibitor (imatinib). The alignment RMSD of the common core region (220 residues) identified by MATT was 2.156 Å; the RMSD of the bound imatinib molecules was 1.736 Å. The MATT alignment is shown for the binding site residue positions analyzed here in Fig. 6.2; the C_α RMSD of the 27 binding site residues shown is 1.169 Å. The aligned computed by MATT is shown in Listing 2.

6.3.2 Binding site residue position selection

All residues having one or more atoms within 5Å of one or more imatinib atoms from the imatinib-bound structures PDB:2PL0 or PDB:3HEC were selected as candidate binding site positions. Candidate positions were eliminated if they corresponded to highly gapped columns in either the PFAM:PKINASE or PFAM:PKINASE_TYR MSAs. After filtering the 27 binding site residue positions shown in Fig. 6.2: `30, 38, 51, 52, 53, 71, 74, 75, 78, 83, 84, 104, 105, 106, 107, 108, 109, 111, 146, 147, 148, 149, 157, 166, 167, 168, 169` (residue numbering according to PDB:3HEC).

```
Core Residues: 220
Core RMSD: 2.156

3HEC:A   R--------------------PTFYRQELNKTIWEVPERYQNLSPVGSGAYG-----SVC    39  (A)
2PL0:A   -GSHMQTQKPQKPWWEDEWEVP--------------RETLKLVERLG-----AGQFGEVW   260  (B)

3HEC:A   AAFDTKTGLRVAVKKLSRPFQSII--HAKRTYRELRLLKHMKHENVIGLLDVFTPARSLE    97  (A)
2PL0:A   MGYYNG-HTKVAVKSLKQ--G---SMSPDAFLAEANLMKQLQHQRLVRLYAVVTQ-----   309  (B)

3HEC:A   EFNDVYLVTHLM-GADLNNIVKC---QKLTDDHVQFLIYQILRGLKYIHSADIIHRDLKP   153  (A)
2PL0:A   --EPIYIITEYMENGSLVDFLKTPSGIKLTINKLLDMAAQIAEGMAFIEERNYIHRDLRA   367  (B)

3HEC:A   SNLAVNEDCELKILDFGLARHTDDEMTGYVA-------------------TRWYRAPEIM   194  (A)
2PL0:A   ANILVSDTLSCKIADF---------------GLARLIEDNEYTAREGAKFPIKWTAPEAI   412  (B)

3HEC:A   LNWMHYNQTVDIWSVGCIMAELLTG-RTLF--PGTDHIDQLKLI--LRLVGTPGAELLKK   249  (A)
2PL0:A   NYGT-FTIKSDVWSFGILLTEIVTHGRIPYPGMTNP--EVIQNLERGYR-----------   458  (B)

3HEC:A   ISSESARNYIQSLTQMPKMNFA-NVFIGANPLAVDLLEKMLVLDSDKRITAAQALAHAYF   308  (A)
2PL0:A   ----------------------MVRPDNCPEELYQLMRLCWKERPEDRPTFDYLRSV-LE   495  (B)

3HEC:A   AQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPP-------------   352  (A)
2PL0:A   D-----------------------------------------FFTATEGQYQPQP   509  (B)
```

Listing 2: Structure-based alignment computed by MATT between TK and non-TK structures PDB:2PL0 and PDB:3HEC, respectively.

6.3.3 Kinase inhibitor affinity annotation labels

The binding affinity (K_d's) for 38 small molecule kinase inhibitor compounds was determined for a set of 317 kinases using an *in vitro* competition binding assay by Karaman et al. 2008 [6]. The 38 inhibitors tested include staurosporine, 1 lipid kinase inhibitor, 15 serine-threonine kinase inhibitors and 21 tyrosine kinase inhibitors. Affinity values were mapped from the Karaman et al. 2008 dataset to kinome structural dataset by mapping the NCBI RefSeq IDs provided by Karaman et al. 2008 to the UniProtKBIDs [83] of the proteins in the structural dataset. 137 of the 208 protein sequences in the structural dataset mapped to the affinity dataset published by Karaman et al. 2008.

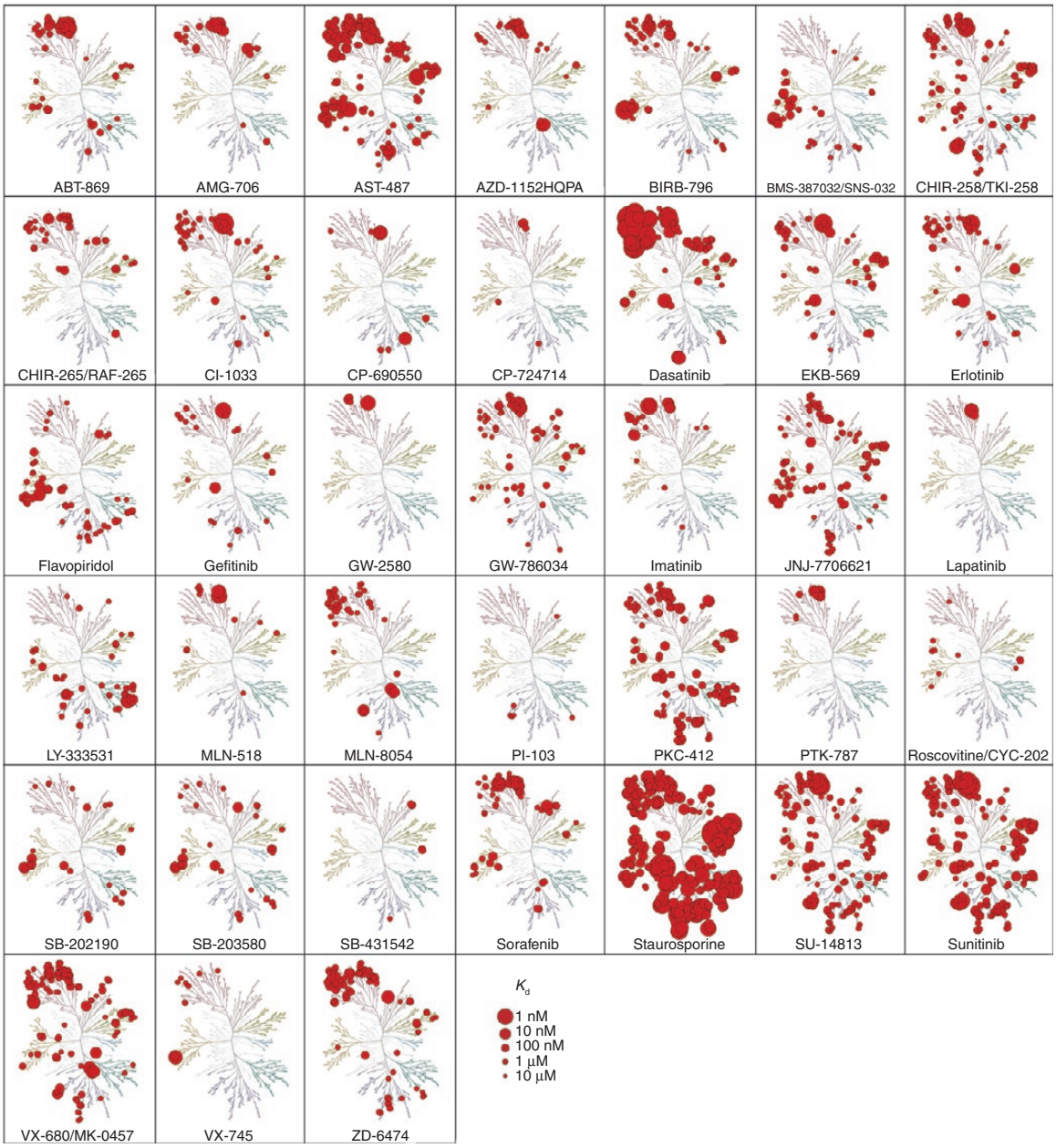


Figure 6.3 : **Kinase affinity dataset.** Kinome affinity maps for the 38 inhibitor dataset of Karaman et al. 2008. Adapted by permission from Macmillan Publishers Ltd: Nature Biotechnology [6], copyright 2008. The kinase dendrogram was adapted and is reproduced with permission from Science (http://www.sciencemag.org) and Cell Signaling Technology, Inc. (http://www.cellsignal.com).

In order to simplify the problem of correlating structural features with binding affinities, the binding affinity (K_d) values were binned into 2 classes (`true`/`false`) by thresholding the affinity values at 10μM(i.e., $<10\mu$M$\rightarrow$ `true`; $\geq 10\mu$M$\rightarrow$ `false`). This cutoff between the two label classes was used consistently across all inhibitors.

6.4 Problem definition

Given the aligned kinase binding site substructures selected in Sec. 6.3.2 and the inhibitor affinity annotation labels as generated in Sec. 6.3.3, the ability of CCORPS to predict the affinity for each inhibitor was assessed using the cross-fold validation approach described in Sec. 4.7. For each of the 38 inhibitor annotation label sets, an independent evaluation of CCORPS was performed. No information was shared between the evaluations in order to validate the predictive ability of CCORPS to identify structural features predictive of the binding ability of each inhibitor independently.

6.5 Prediction performance

For each of the 38 inhibitors included in the affinity dataset, CCORPS was used to predict the set of kinases able to bind to that inhibitor. The performance of CCORPS was assessed for each inhibitor, independently, by computing the Receiver Operator Characteristic (ROC) curve [91] for the set of predictions, which evaluates the sensitivity (# true positives / (# true positives + # false negatives)) at each specificity (# true negatives / (# true negatives + # false positives)) value. The ROC curves for the predictor constructed by CCORPS are shown in Fig. 6.4 for each inhibitor and the Area Under Curve (AUC) [91] for each ROC curve is listed in Table 6.1. Additionally, the Precision-Recall (PR) curve [91] for each inhibitor can

be found in Fig. 6.5. The PR curve plots the precision (# true positives / (# true positives + # false positives)) versus the recall (equivalent to sensitivity).

In order to make a direct comparison of the performance of CCORPS to the work of Jackson et al. 2009, another performance measure, the enrichment factor, was also computed per inhibitor tested. The enrichment factor of the top 5% most highly ranked `true` affinity predictions (for a given inhibitor) can be calculated as follows:

$$\mathrm{E}_{5\%} = \frac{A_{\text{top }5\%}/N_{\text{top }5\%}}{A_{\text{total}}/N_{\text{total}}} \tag{6.1}$$

where $A_{\text{top }5\%}$ is the number of structures with known affinity for a given inhibitor (# *actives*) in the top 5% of most confident predictions as ranked by CCORPS, $N_{\text{top }5\%}$ is the total number of structures in the top 5%, A_{total} is the total number of active structures in the dataset and N_{total} is total number of structures in the dataset. The enrichment factor at 5% ($\mathrm{E}_{5\%}$) for each inhibitor is shown in Table 6.1 and where available, the corresponding $\mathrm{E}_{5\%}$ values from Jackson et al. 2009 are listed alongside. It should be noted that the $\mathrm{E}_{5\%}$ values are *not directly* comparable between CCORPS and Jackson et al. 2009 as listed in Table 6.1, due to the fact that the maximum possible enrichment (E_{max}) for a given inhibitor is dataset-dependent, and the dataset presented in this work is larger both in number of structures compared and the number of per-inhibitor affinity annotations.

As shown in Table. 6.1, CCORPS achieves high predictive performance across the 38 inhibitors tested. As quantified by $\mathrm{E}_{5\%}$, CCORPS achieved perfect or near-perfect predictive ability for 12 of the 38 inhibitors (AST-487, BMS-387032, CHIR-258, CHIR-265, EKB-569, GW-786034, MLN-8054, roscovitine, SB-202190, sorafenib, sunitinib and ZD-6474), while only exhibiting overall poor performance for one inhibitor, GW-2580. CCORPS was unable to identify predictive structural features consistent among kinases capable of bind-

ing GW-2580. However, features consistent among kinases *unable* to bind GW-2580 were identified, resulting in CCORPS (incorrectly) predicting the `false` label for all kinases with respect to GW-2580 binding. To summarize, CCORPS is demonstrated to consistently meet or exceed the predictive ability of the method by Jackson et al. 2009 as also shown in Table 6.1.

6.6 Highly predictive clusters

In the process of identifying structural features within the kinome that are predictive of affinity for a particular inhibitor, CCORPS computes the substructure clustering for all possible 3-position subsets of the original 27-position binding site definition (see Sec. 6.3.2). One of the 2925 substructure clusterings computed for the kinome is shown in Fig. 6.6**(b)**. Each point within Fig. 6.6**(b)** represents the feature vector, as computed by CCORPS, for a single 3-position substructure; that is, each of the 1958 substructures within the kinase structure dataset is shown. As demonstrated in Fig. 6.6**(c)**, where one randomly selected representative substructure is shown for each of the 21 clusters identified by CCORPS, both the geometry and residue types vary significantly among the kinases.

Because of the structural and chemical variance exhibited by kinases at the 3-position substructure shown in Fig. 6.6**(c)**, several Highly Predictive Clusters (HPCs) can be identified in Fig. 6.6**(b)**. These HPCs are clusters consisting of substructures all sharing the same affinity annotation label (either all `true` or all `false`). As discussed in Sec. 4.5.1, affinity annotation labels are predicted by tallying the number of times an unannotated structure is co-clustered into an HPC for each possible label (`true` and `false` here) and then applying the SVM-based decision boundary, as was detailed in Sec. 4.6. Note that the final prediction made by CCORPS for any particular structure is based upon the HPCs identified across *all*

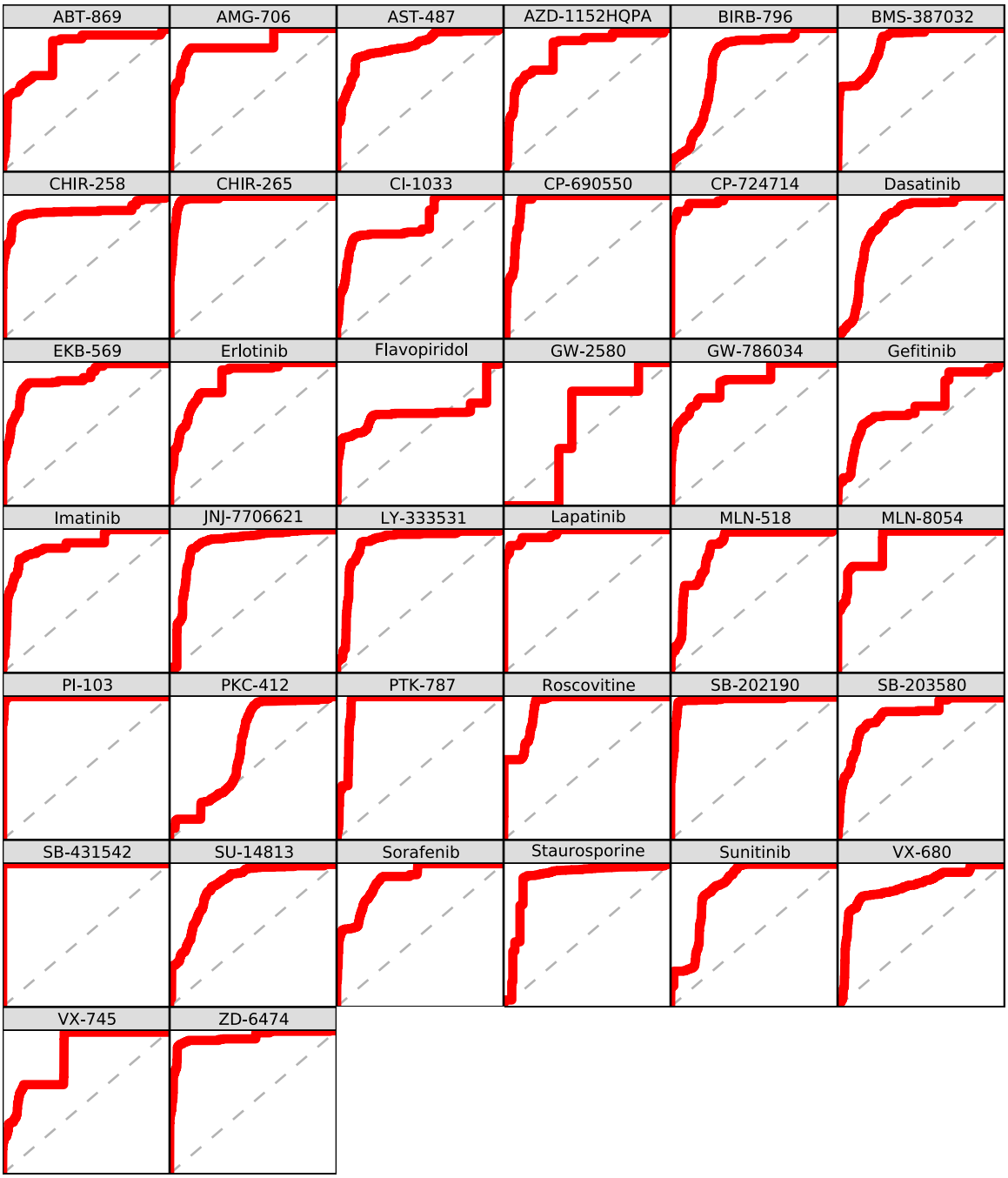


Figure 6.4 : **Per drug Receiver Operator Characteristic (ROC) curves**. The *x*- and *y*-axis plot (1-specificity) and sensitivity, respectively, both ranging from 0 to 1. The Area Under Curve (AUC_{ROC}) as well as the $E_{5\%}$ per drug can be found in in Table 6.1. As shown above, CCORPS is able to construct a near-perfect classifier for several drugs, such as PI-103, SB-431542. The classifiers constructed for some inhibitors, such as flavopiridol, are able to achieve high precision, but only at low sensitivities (recalls), as further illustrated by the PR curves in Fig. 6.5.

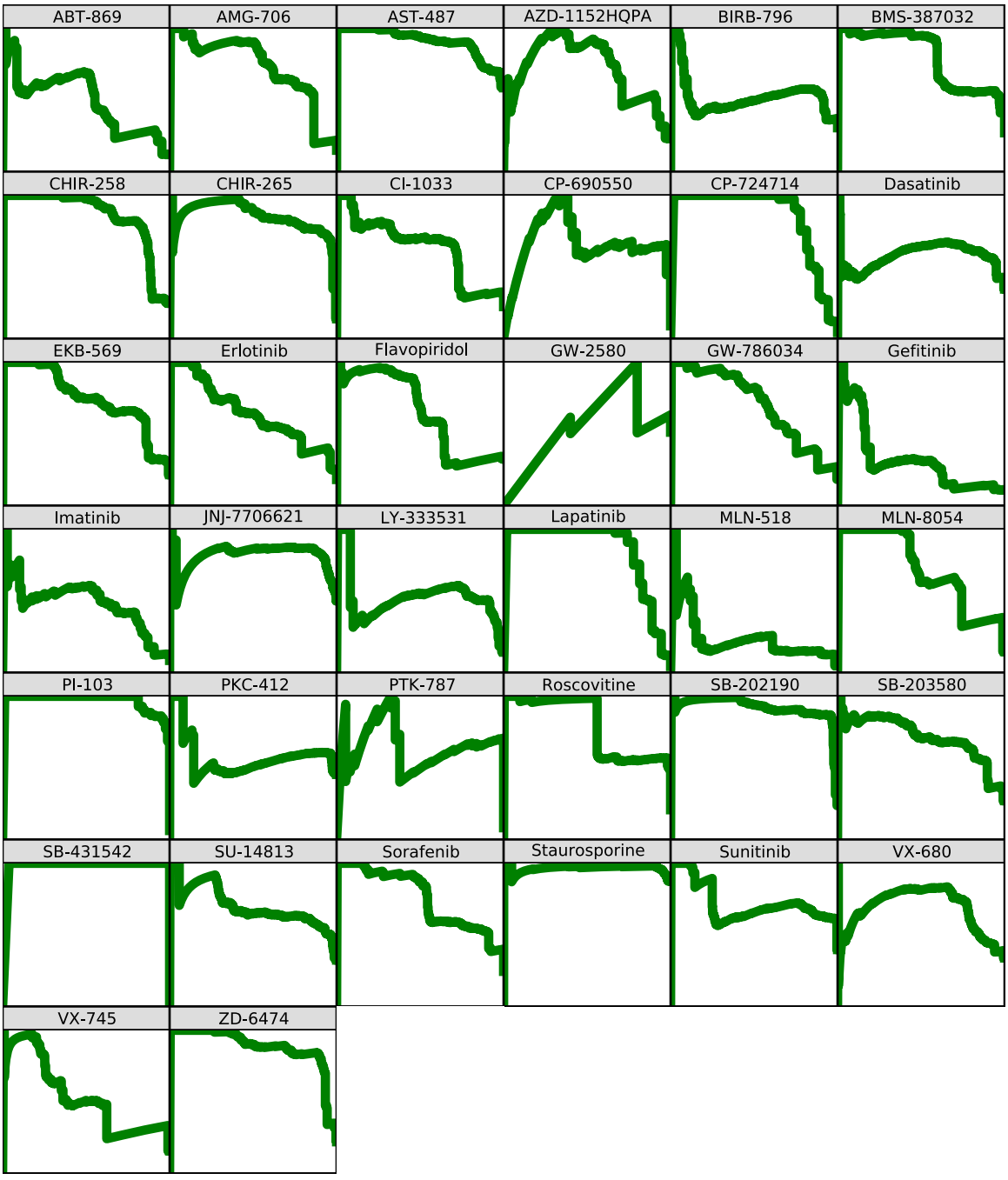


Figure 6.5 : **Per drug Precision-Recall (PR) curves.** The *x*- and *y*-axis plot the recall and precision, respectively, both ranging from 0 to 1. The Area Under Curve (AUC_{PR}) per drug can be found in Table 6.1. As shown above, CCORPS is demonstrated to have very high precision across a wide range of inhibitors when tested for targets spanning the kinome.

Inhibitor	CCORPS AUC_{ROC}	CCORPS AUC_{PR}	CCORPS $E_{5\%}$	CCORPS E_{max}	Jackson et al. 2009 $E_{5\%}$	Jackson et al. 2009 E_{max}
ABT-869	0.84	0.49	5.14	(8.43)		
AMG-706	0.89	0.73	5.97	(6.71)		
AST-487	0.87	0.91	1.71	(1.71)		
AZD-1152HQPA	0.84	0.45	3.71	(6.78)		
BIRB-796	0.77	0.53	1.99	(3.27)	3.65	(3.98)
BMS-387032	0.90	0.80	3.52	(3.69)		
CHIR-258	0.89	0.86	4.05	(4.05)		
CHIR-265	0.97	0.85	7.01	(7.24)		
CI-1033	0.82	0.64	3.57	(4.48)		
CP-690550	0.94	0.22	8.21	(32.85)		
CP-724714	0.97	0.85	16.31	(23.72)		
Dasatinib	0.81	0.58	1.31	(2.89)		
EKB-569	0.89	0.75	4.57	(4.64)		
Erlotinib	0.88	0.65	4.64	(5.50)	6.89	(9.19)
Flavopiridol	0.65	0.65	2.90	(3.09)		
GW-2580	0.54	0.00	0.00	(256.20)		
GW-786034	0.86	0.70	5.20	(5.45)		
Gefitinib	0.70	0.32	3.63	(9.28)		
Imatinib	0.89	0.48	6.49	(11.86)	2.99,5.98	(11.95)
JNJ-7706621	0.89	0.82	1.13	(2.00)		
LY-333531	0.90	0.53	2.53	(7.04)		
Lapatinib	0.97	0.85	16.31	(23.72)	0.00	(19.92)
MLN-518	0.87	0.24	3.73	(21.71)		
MLN-8054	0.91	0.72	6.85	(6.85)		
PI-103	1.00	0.97	16.01	(16.01)		
PKC-412	0.66	0.58	1.62	(2.20)		
PTK-787	0.95	0.23	7.57	(34.62)		
Roscovitine	0.93	0.79	4.67	(4.82)	2.81	(2.81)
SB-202190	0.97	0.92	4.15	(4.21)		
SB-203580	0.87	0.71	3.11	(3.69)	5.43	(5.43)
SB-431542	1.00	0.98	20.02	(49.27)		
SU-14813	0.85	0.70	2.60	(3.09)		
Sorafenib	0.89	0.75	3.82	(4.08)		
Staurosporine	0.89	0.97	1.00	(1.15)		
Sunitinib	0.83	0.72	2.48	(2.52)		
VX-680	0.83	0.68	1.66	(2.95)		
VX-745	0.83	0.54	5.05	(6.34)		
ZD-6474	0.95	0.88	4.53	(4.53)		

Table 6.1 : **Affinity prediction performance of CCORPS for the kinase inhibitors.** For each of the 38 inhibitors in the affinity dataset of Karaman et al. 2008, the prediction performance of CCORPS is shown above for the performance metrics discussed in Sec. 6.5. The performance of the Jackson et al. 2009 method is shown alongside that of CCORPS for the subset of inhibitors tested by both methods. Note that for imatinib, two $E_{5\%}$ values are provided by Jackson et al. 2009 because each value is derived by selecting a different reference structure as discussed in Sec. 6.2.

of the 2925 substructure clusterings, not just the single clustering illustrated in Fig. 6.6.

Several informative observations regarding kinase structural diversity and its association to inhibitor binding affinity can be made by further examination of the substructure clustering shown in Fig. 6.6**(b)**. Immediately upon examination of the substructure clustering it can be noted that multiple distinct clusters of kinases exist. This observation alone indicates that the 3-position substructure that resulted in this clustering is highly diverse among kinase binding sites. Conversely, the presence of a single large cluster would indicate that the 3-position substructure was structurally conserved, exhibiting little variance across the kinome; indeed instances of clusterings with a single dominating cluster were also observed for some 3-position subsets. As demonstrated in Fig. 6.6**(c)**, where one randomly selected representative substructure is shown for each of the 21 clusters identified by CCORPS, both the geometry and residue types vary significantly for this 3-position subset.

By incorporating the affinity annotation labels for a particular inhibitor, further observations can be made about the association between the 3-position substructure shown in Fig. 6.6**(a)** and the kinases capable of binding that inhibitor. For example, mapping the affinity annotation labels for the inhibitor flavopiridol onto the substructure clustering (Fig. 6.6**(d)**) reveals that some of the clusters consist purely of only a single annotation label while others are a mixture of labels. In Fig. 6.6**(d)**, kinases capable of binding flavopiridol are colored red (`true` label), kinases incapable of binding flavopiridol are colored gray (`false` label) and kinases lacking affinity annotation are colored white (undefined label). As illustrated in Fig. 4.2, HPCs are clusters that consist of a single annotation label. As shown in Fig. 6.6**(d)**, multiple clusters of purely `true` labels exist and are considered to be HPCs by CCORPS.

The existence of `true`-only clusters indicates that the 3-positions shown in Fig. 6.6**(a)** are a distinguishing structural feature for identifying kinases that bind flavopiridol. More

interestingly, however, is the fact that multiple, structurally distinct versions of the same 3-position substructure exist for different kinases that all are capable of binding flavopiridol. This result is significant because it indicates that different kinases have different structural motifs that are associated with binding flavopiridol, as opposed to a single, shared structural motif across all flavopiridol-binding kinases.

Furthermore, the existence of clusters containing only kinases *incapable* of binding flavopiridol can also be observed in Fig. 6.6**(d)**. These HPCs are also informative because they identify particular structural versions of the 3-position substructure in Fig. 6.6**(a)** that are all incapable of binding flavopiridol. Finally, clusters consisting of a mixture of kinases that are both capable and incapable of binding flavopiridol can be identified in Fig. 6.6**(d)**. For kinases in these clusters, the 3-position substructure is not a distinguishing feature of flavopiridol-binding ability.

Finally, while flavopiridol is discussed in detail here, the same analysis was computed by CCORPS for each of the 38 different inhibitors within the affinity dataset. For each of the inhibitors, the affinity labels can be mapped separately onto the same substructure clustering as shown in Fig. 6.7. However, it should be noted that no information is shared between the results for different inhibitors in this work; that is, each inhibitor is computed in a fully separate CCORPS computation (the substructure clusterings do not vary, just the annotation labels).

Examination of the affinity-annotated substructure clusterings shown in Fig. 6.7 reveals that the set of clusters which are HPCs varies greatly depending on the inhibitor considered. While the flavopiridol-annotated substructure clustering contains multiple HPCs for both `true` and `false` labels, the correspondingly annotated clustering for other inhibitors, such as VX-745, PI-103 and imatinib, contain only `false` HPCs. This result demonstrates that the substructures that are informative of inhibitor binding are inherently inhibitor-specific.

That is, a subset of binding site positions that are predictive for one inhibitor are not necessarily predictive for another inhibitor.

6.7 Phylogenetically diverse HPCs

Numerous instances of cross-family affinity for both type I and II kinase inhibitors have been identified, as is clearly illustrated by the kinome affinity maps created by Karaman et al. 2008 (Fig. 6.3) using their kinome-wide affinity screening dataset. It is important to identify structural features shared among phylogenetically diverse kinases that share affinity for a particular inhibitor, because they provide a basis for reasoning about inhibitor cross-reactivity. Furthermore, by identifying these shared structural features, it may be possible to rationally reengineer the specificity of inhibitors by avoiding the targeting of these features, since they are not unique to the intended kinase target. In order to identify the number of instances of cross-family structural features that can be associated with specific inhibitor binding, the distribution of substructure clusters across all 3-position subsets was analyzed.

Each individual cluster, across all 2925 clusterings and all 38 inhibitors, was evaluated to calculate the SS-NR-purity of both affinity labels and family-level phylogenetic labels. For example, a cluster containing 3 distinct kinase sequences with affinity labels {`true`, `false`, `true` } and family labels {AGC, CAMK, TK } would have an affinity SS-NR-purity of 0.66 and a phylogenetic SS-NR-purity of 0.33. By plotting the affinity and phylogenetic SS-NR-purity scores of each cluster (separately for each inhibitor) as shown in Fig. 6.8, the distribution of clusters across the spectrum of possible scores can be evaluated. As can be seen in Fig. 6.8, many phylogenetically diverse clusters of 100% affinity purity labels exist. This result indicates that shared structural features across family-level boundaries within

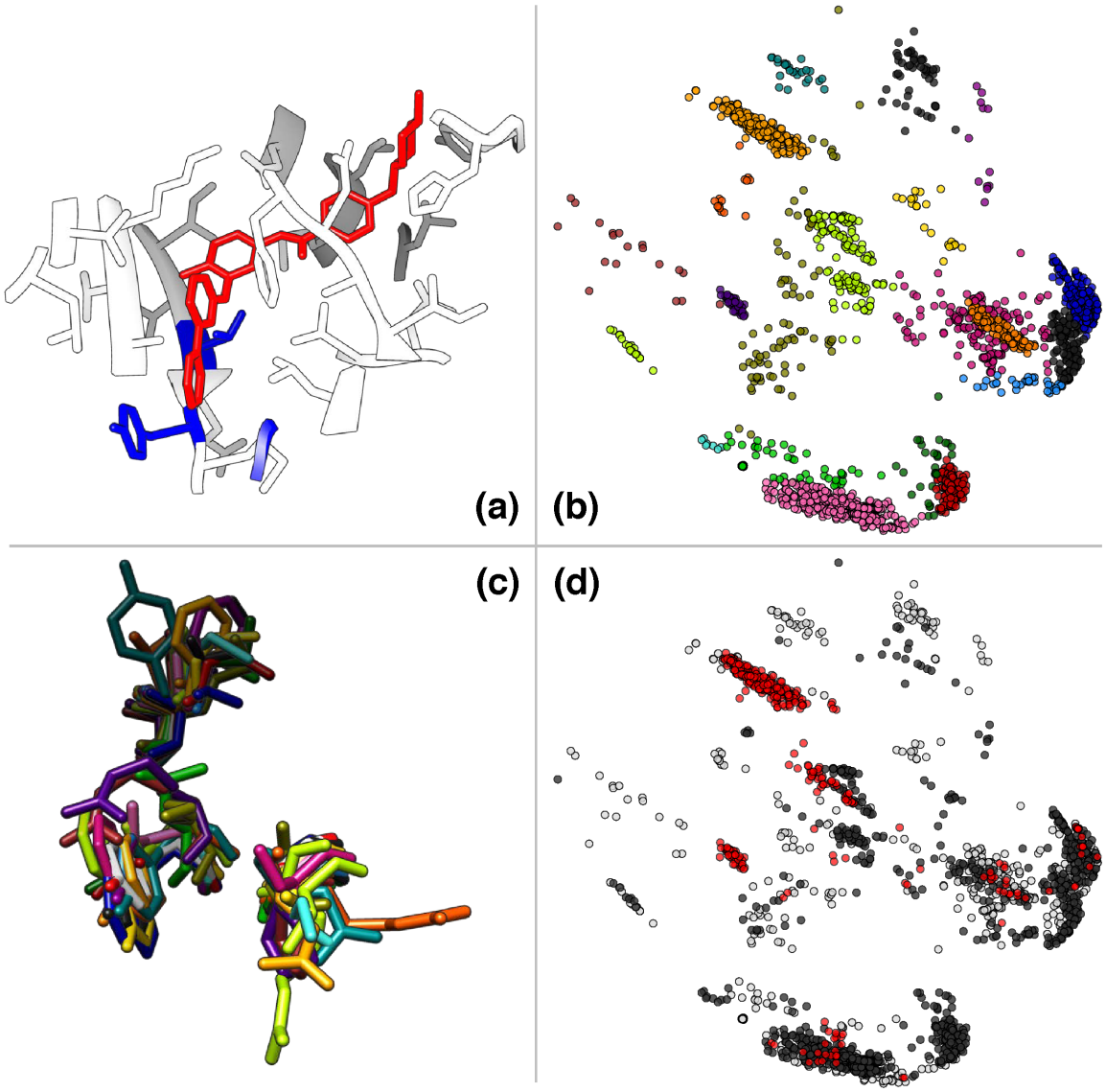


Figure 6.6 : **Highly predictive clusters**. **(a)** Structure of lck (PDB:2PL0) with 3-position substructure shown in blue stick representation (Thr-316, Tyr-318, Gly-322) and bound imatinib molecule in red. **(b)** Substructure clustering computed by CCORPS when comparing the 3-positions shown in **(a)** across the entire 1958 structure dataset. Each point in the clustering represents a single 3-residue substructure. The red and black coloring of each point indicates `true` and `false` affinity labels for flavopiridol, respectively, while white indicates substructures lacking affinity annotations. **(c)** Aligned 3-residue substructure representatives from each of the 21 clusters identified by CCORPS for the 3-position subset shown in **(a)**. The color of each substructure corresponds to its cluster assignment. **(d)** The same substructure clustering shown in **(b)** but relabeled to indicate the cluster membership of each substructure (21 clusters in total are shown). It should be noted that only one of the 2925 clusterings computed by CCORPS is shown above, with only one of the 38 inhibitor affinity labelings shown.

the kinome exist and can be correlated with shared binding affinity for particular inhibitors. As further demonstrated by the corresponding distributions for all 38 inhibitors in Fig. 6.9, such shared structural similarity is not rare.

In order to build intuition for interpreting the cluster distributions, the cluster distribution for VX-680 (Fig. 6.8) is first examined in more detail because it is representative of the distribution for many of the other inhibitors. As listed in Table 6.2, 23,495 clusters were identified by CCORPS that have ≥ 0.5 SS-NR-purity in the `true` label for VX-680 (hereafter referred to as `true`-majority clusters). Only these `true`-majority clusters are plotted in the cluster distribution shown in Fig. 6.8, meaning the minimum "affinity purity" displayed in Fig. 6.8 is 0.5 by definition (because only 2 different affinity labels exist, `true` and `false`).

As can be seen in Fig. 6.8, the vast majority of clusters identified by CCORPS have low affinity SS-NR-purity as well as low phylogenetic SS-NR-purity. This is to be expected because highly conserved portions of the kinase ATP binding site are known to exist. Structural features that consist of conserved residue positions will be common to many kinases from different families due to the fact that these positions are so heavily conserved, which explains the low phylogenetic SS-NR-purity of these clusters. Furthermore, these conserved features are unlikely to be correlated with the affinity for a particular inhibitor because most inhibitors have been engineered to not have broad cross-reactivity across the kinome. However, staurosporine is known to have broad cross reactivity due to its interaction with highly conserved binding site features; the cluster distribution corresponding to staurosporine (Fig. 6.9) is markedly different from the other inhibitors with most clusters having high affinity SS-NR-purity across a range of phylogenetic SS-NR-purity values.

Examination of the extremes of the VX-680 cluster distribution reveals further insights into the frequency of structural similar features among kinases with different degrees of sequence similarity. Clusters having a phylogenetic SS-NR-purity of 1.0 (i.e., all proteins

belong to the same family) but having low affinity SS-NR-purity exist, and for VX-680 in particular, 276 such clusters were identified by CCORPS. This observation is interesting because it illustrates that kinases sharing sequence similarity (relative to kinases outside the family) have multiple common structural features that are not informative of the ability of these kinases to bind VX-680. Because CCORPS only incorporates clusters with high affinity SS-NR-purity (i.e., HPCs), these conserved structural features that are not indicative of VX-680 binding are ignored by CCORPS when predicting affinity for unannotated kinases. This observation can also be made for each of the other inhibitors as shown in Fig. 6.9.

Another interesting extreme of the VX-680 cluster distribution to examine is the existence of HPCs that are phylogenetically diverse. The HPCs selected by CCORPS correspond to the right-most column of points in Fig. 6.8; these clusters all have an affinity SS-NR-purity of 1.0 for VX-680 and therefore contain *only* structures with known VX-680 affinity. As can be noted in Fig. 6.8, HPCs exist at a range of phylogenetic SS-NR-purity values. CCORPS identified a total of 2707 HPCs for VX-680, and 1786 (66%) of these HPCs contain proteins belonging to two or more distinct kinase families. This result demonstrates that CCORPS is capable of identifying cross-family structural features that are associated with VX-680 binding. Furthermore, this result is not unique to VX-680. As shown in Fig. 6.9 and tabulated in Table 6.2, cross-family structural features associated with inhibitor binding were identified for all of the inhibitors tested with the exception of GW-2580, for which no `true`-majority HPCs were identified.

Examination of the cluster distributions across each of the inhibitors reveals a wide range of observations. While many inhibitors have a cluster distribution similar to that of VX-680, for some inhibitors CCORPS identified relatively fewer `true`-majority clusters. For example, only 133 clusters with affinity SS-NR-purity >0.5 were identified by CCORPS for SB-431542 and all of these happen to be HPCs. However, even among this relatively

low number of HPCs, 69 (52%) of the clusters contain kinases from two or more families. As demonstrated by the corresponding distributions for all 38 inhibitors in Fig. 6.9, such shared structural similarity is not rare.

6.8 Conclusion

The high degree of ATP binding site similarity shared across the protein kinases has made them a difficult target for which to design highly selective inhibitors. However, by identifying the patterns of local structural similarity among binding sites at the kinome scale, potential off-target interactions may be able to be identified at earlier stages of pharmaceutical development and compensated for by further inhibitor modification.

As was demonstrated here, CCORPS is capable incorporating all of the available protein kinase structure data, so as to operate at the kinome scale, and then uses this data to construct highly accurate predictors of kinase affinity for a variety of different small molecule inhibitors. While CCORPS relies upon the aggregation of structural similarity that coincides with affinity similarity to build predictors, the individual instances may be informative in and of themselves. Further analysis of the vast number of structurally similar features shared among phylogenetically distant kinases may provide additional insights into the structural mechanisms of inhibitor recognition occurring across the kinome. Structural features that are found to be unique to one or a small number of chosen kinases may provide the starting point for designing highly specific inhibitor interactions and therefore highly selective protein kinase inhibitors.

Inhibitor	# true-HPCs	# ≥2 Families
ABT-869	345	249
AMG-706	274	202
AST-487	2415	1955
AZD-1152HQPA	506	374
BIRB-796	893	730
BMS-387032	728	447
CHIR-258	1577	800
CHIR-265	242	184
CI-1033	1247	704
CP-690550	11	5
CP-724714	115	89
Dasatinib	1848	1193
EKB-569	1133	684
Erlotinib	596	456
Flavopiridol	921	481
GW-2580	0	0
GW-786034	1443	809
Gefitinib	203	169
Imatinib	57	45
JNJ-7706621	4087	2761
LY-333531	634	314
Lapatinib	115	89
MLN-518	92	70
MLN-8054	435	301
PI-103	182	69
PKC-412	3419	2368
PTK-787	7	6
Roscovitine	593	335
SB-202190	644	513
SB-203580	738	546
SB-431542	133	69
SU-14813	4415	3116
Sorafenib	561	405
Staurosporine	17098	14802
Sunitinib	5525	4077
VX-680	2707	1786
VX-745	189	151
ZD-6474	1059	610

Table 6.2 : **Phylogenetically diverse HPC statistics per inhibitor.** For each inhibitor, the total number of true-HPCs (column "# true-HPCs") is shown. The subset of true-HPCs that consist of proteins from two or more of the kinase families defined by Manning et al. [4] (column "# ≥2 families") are also shown. The multitude of true-HPCs that include proteins from distinct families of the kinome can be noted by the relatively large percentage (73% overall across all inhibitors) of HPCs that span families.

Figure 6.7 : **Affinity annotation labeling for all 38 inhibitors.** The substructure clustering computed for the same 3 positions examined in Fig. 6.6 is relabeled above for each of the 38 inhibitors included in the dataset. In each cell above, red and black indicate the `true` and `false` affinity labels, respectively, for each inhibitor, while white indicates a lack of annotation. As can be noted by comparing the distribution of red points across the different inhibitors, for most inhibitors, the kinase proteins capable of binding to them are not distributed in a single cluster, indicating structurally diverse features exist among the kinases selected by each inhibitor.

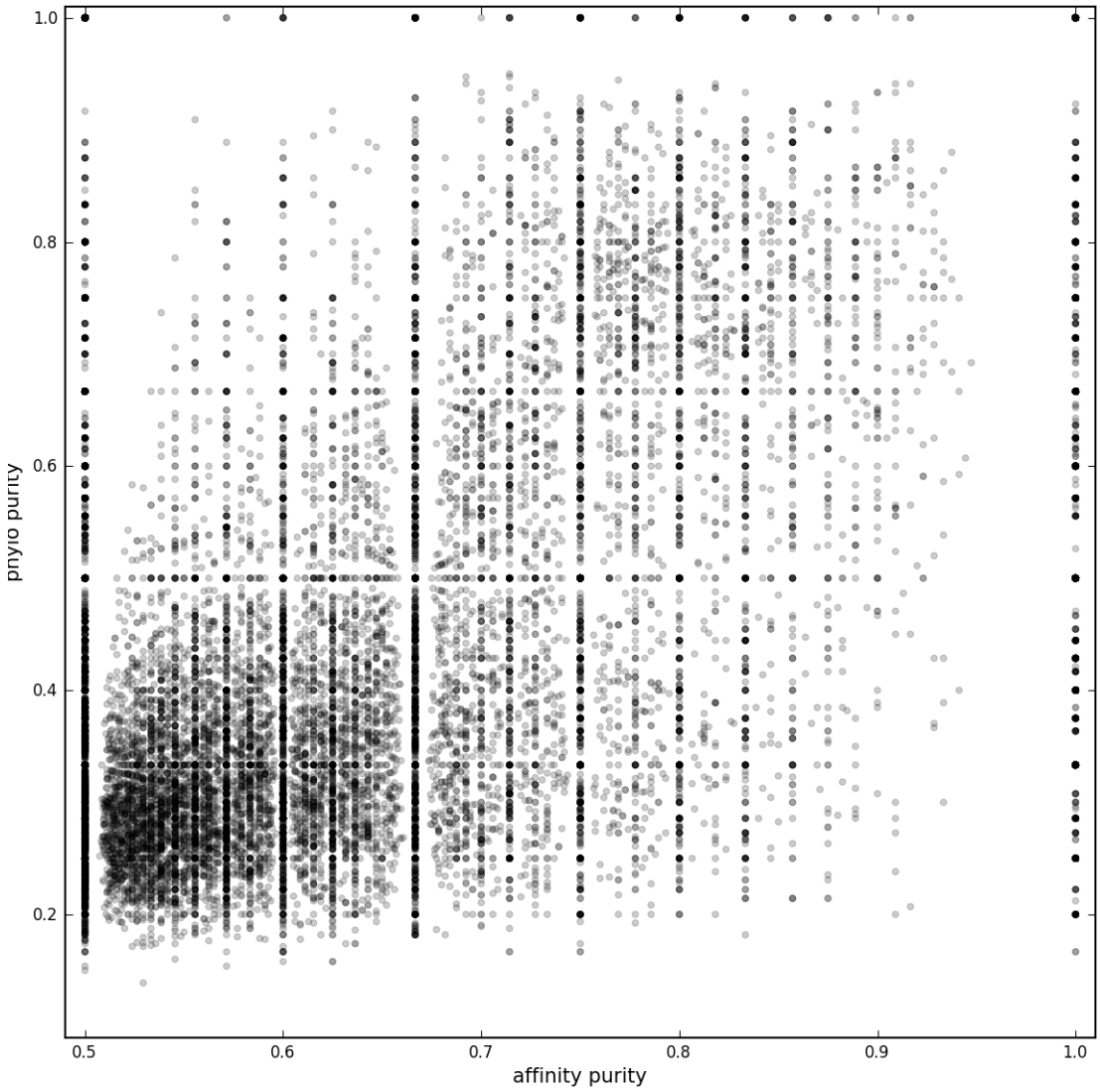


Figure 6.8 : **Distribution of phylogenetic and affinity SS-NR-purity cluster scores for VX-680.** Each point in the scatter plot above marks the SS-NR-purity for the drug affinity `true` label on the *x*-axis and the phylogenetic label SS-NR-purity on the *y*-axis. For example, a point above located at the coordinates `(1.0, 0.2)` denotes a cluster that is 100% pure in the `true` drug affinity label (for VX-680 in this case) but is only 20% pure in the most common phylogenetic label present; that is, this cluster indicates one instance of structural similarity among phylogenetically diverse proteins that is also coincides with having affinity for VX-680. Conversely, a point at the coordinates `(0.5, 1.0)` indicates a cluster that contains only structures from one phylogenetic (family-level) branch but contains an equal proportion of `true` and `false` affinity labels; that is, a case where structurally similar, closely related (phylogenetically) structures have different affinities for VX-680.

Figure 6.9 : **Distribution of phylogenetic and affinity SS-NR-purity cluster scores for all drugs**. As can be seen in the case of drugs such as imatinib and lapatinib, very few clusters that have a majority of `true` labels were identified, yet clusters of phylogenetically diverse structures all having `true` labels can be identified. Staurosporine exhibits a reflected distribution relative to the other drugs, because due the nature of its non-selectivity across the kinome, instances of phylogenetically distant structures that exhibit Staurosporine affinity are common. Refer to Fig. 6.8 for additional details.

Chapter 7

Conclusion

This thesis introduces two fundamental approaches for learning structure-function relationships among large protein structure datasets.

The FASST Framework demonstrates a highly modular unsupervised learning approach to identifying the variability of local structural features among large numbers of structures for the selection of unique structural conformations and the detection of outlier structures. Additionally, the interactive FASST Live visualization and analysis tool provides a platform for the broad dissemination of such techniques into the research community.

The supervised learning approach implemented by CCORPS provides an approach to automatically identify the structural features of large families of proteins that give rise to different sets of annotation labels, such as ligand binding affinity and functional diversification. The ability of CCORPS to accurately predict the affinity of protein kinase inhibitors across the human kinome for the 38 compound dataset was demonstrated. Furthermore, the generality of the approach implemented by CCORPS was also demonstrated by predicting the enzymatic function class of proteins from 48 different functionally diverse families.

Together, these approaches provide a principled mechanism for the incorporation of

all available structural data. Instead of considering the abundance of alternative structures for many proteins a hindrance that necessitates careful filtering and removal, the methods presented here harness structural overrepresentation to provide an unbiased representation of the true variety of protein structure datasets.

As the quantity of both protein structure data and protein functional annotations continue to grow, the learning methods introduced here will become increasingly necessary in order to undertake comparative structural studies at the level of whole protein families.

Future work may take advantage of the additional structural data available from protein-ligand co-crystal structures. For example, the kinase inhibitor experiments presented in this thesis could be further enhanced by additionally considering the 3-dimensional conformation of the bound inhibitors and the specific interactions (e.g., hydrogen bonding) being made to the kinase binding site residue positions. Furthermore, many kinase inhibitors have common chemical groups and identifying the connection between shared inhibitor chemical groups and shared inhibition of specific sets of kinases would further increase our current understanding of kinase inhibition and may lead to the development of novel, highly-selective kinase inhibitors.

Bibliography

[1] D. H. Bryant, M. Moll, B. Y. Chen, V. Y. Fofanov, and L. E. Kavraki, "Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction," *BMC Bioinformatics*, vol. 11, p. 242, 2010.

[2] M. Moll, D. H. Bryant, and L. E. Kavraki, "The LabelHash algorithm for substructure matching," *BMC Bioinformatics*, vol. 11, p. 555, 2010.

[3] E. C. Meng, B. J. Polacco, and P. C. Babbitt, "Superfamily active site templates.," *Proteins*, vol. 55, pp. 962–976, Jun 2004.

[4] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, pp. 1912–34, Dec 2002.

[5] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLoS Comput Biol*, vol. 4, p. e10, Jan 2008.

[6] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar, "A quantitative analysis of kinase inhibitor selectivity," *Nat Biotechnol*, vol. 26, pp. 127–32, Jan 2008.

[7] T. Hunter, "The role of tyrosine phosphorylation in cell growth and disease," *Harvey Lect*, vol. 94, pp. 81–119, 1998.

[8] T. Zarubin and J. Han, "Activation and signaling of the p38 map kinase pathway," *Cell Res*, vol. 15, pp. 11–8, Jan 2005.

[9] P. Cohen, "Protein kinases–the major drug targets of the twenty-first century?," *Nat Rev Drug Discov*, vol. 1, pp. 309–15, Apr 2002.

[10] Y. Liu and N. S. Gray, "Rational design of inhibitors that bind to inactive kinase conformations," *Nat Chem Biol*, vol. 2, pp. 358–64, Jul 2006.

[11] D. Huang, T. Zhou, K. Lafleur, C. Nevado, and A. Caflisch, "Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis," *Bioinformatics*, vol. 26, pp. 198–204, Jan 2010.

[12] P. F. Gherardini, M. N. Wass, M. Helmer-Citterich, and M. J. E. Sternberg, "Convergent evolution of enzyme active sites is not a rare phenomenon," *Journal of Molecular Biology*, vol. 372, no. 3, pp. 817–845, 2007.

[13] D. Rognan, "Chemogenomic approaches to rational drug design.," *British Journal of Pharmacology*, vol. 152, pp. 38–52, Sep 2007.

[14] P. F. Gherardini and M. Helmer-Citterich, "Structure-based function prediction: approaches and applications," *Brief Funct Genomic Proteomic*, vol. 7, pp. 291–302, Jul 2008.

[15] R. Minai, Y. Matsuo, H. Onuki, and H. Hirota, "Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions.," *Proteins*, vol. 72, no. 1, pp. 367–381, 2008 Jul.

[16] Y. Y. Tseng and W.-H. Li, "Classification of protein functional surfaces using structural characteristics," *Proc Natl Acad Sci U S A*, vol. 109, pp. 1170–5, Jan 2012.

[17] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[18] M. Hult, N. Shafqat, B. Elleby, D. Mitschke, S. Svensson, M. Forsgren, T. Barf, J. Vallgarda, L. Abrahmsen, and U. Oppermann, "Active site variability of type 1 11beta-hydroxysteroid dehydrogenase revealed by selective inhibitors and cross-species comparisons.," *Molecular and Cellular Endocrinology*, vol. 248, pp. 26–33, Mar 2006.

[19] G. Petsko and D. Ringe, *Protein structure and function*. Sinauer Associates Inc, 2004.

[20] R. Duda, P. Hart, and D. Stork, *Pattern classification*. John Wiley & Sons, 2nd ed., 2001.

[21] R. Kolodny, D. Petrey, and B. Honig, "Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 393–398, 2006.

[22] Z. Zhang and M. G. Grigorov, "Similarity networks of protein binding sites.," *Proteins*, vol. 62, no. 2, pp. 470–478, 2006 Feb 1.

[23] L. Xie and P. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proceedings of the National Academy of Sciences*, vol. 105, no. 14, p. 5441, 2008.

[24] S. C.-H. Pegg, S. D. Brown, S. Ojha, J. Seffernick, E. C. Meng, J. H. Morris, P. J. Chang, C. C. Huang, T. E. Ferrin, and P. C. Babbitt, "Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database.," *Biochemistry*, vol. 45, pp. 2545–2555, Feb 2006.

[25] E. Webb, "International union of biochemistry and molecular biology. nomenclature committee. enzyme nomenclature, 1992: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes," 1992.

[26] E. C. Meng and P. C. Babbitt, "Topological variation in the evolution of new reactions in functionally diverse enzyme superfamilies," *Curr Opin Struct Biol*, vol. 21, pp. 391–7, Jun 2011.

[27] T. Klabunde, "Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.," *British Journal of Pharmacology*, vol. 152, pp. 5–7, Sep 2007.

[28] A. L. Bowman, M. G. Lerner, and H. A. Carlson, "Protein flexibility and species specificity in structure-based drug discovery: dihydrofolate reductase as a test system.," *Journal of the American Chemical Society*, vol. 129, pp. 3634–3640, Mar 2007.

[29] S. Tyagi and J. Pleiss, "Biochemical profiling in silico–predicting substrate specificities of large enzyme families.," *J Biotechnol*, vol. 124, pp. 108–116, Jun 2006.

[30] R. Finn, J. Tate, J. Mistry, P. Coggill, S. Sammut, *et al.*, "The Pfam protein family database," *Nucleic Acid Research. In press. Find this article online*, 2008.

[31] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH–a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1108, 1997.

[32] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant, "Cdd: a conserved domain database for the functional annotation of proteins," *Nucleic Acids Res*, vol. 39, pp. D225–9, Jan 2011.

[33] O. C. Redfern, B. H. Dessailly, T. J. Dallman, I. Sillitoe, and C. A. Orengo, "Flora: a novel method to predict protein function from structure in diverse superfamilies.," *PLoS Comput Biol*, vol. 5, p. e1000485, Aug 2009.

[34] R. R. Thangudu, M. Tyagi, B. A. Shoemaker, S. H. Bryant, A. R. Panchenko, and T. Madej, "Knowledge-based annotation of small molecule binding sites in proteins," *BMC Bioinformatics*, vol. 11, p. 365, 2010.

[35] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, no. 5275, pp. 595–603, 1996.

[36] A. R. Kinjo and H. Nakamura, "Comprehensive structural classification of ligand-binding motifs in proteins.," *Structure*, vol. 17, no. 2, pp. 234–246, 2009 Feb 13.

[37] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[38] R. V. Spriggs, P. J. Artymiuk, and P. Willett, "Searching for patterns of amino acids in 3d protein structures," *J Chem Inf Comput Sci*, vol. 43, no. 2, pp. 412–21, 2003.

[39] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among

proteins independent of sequence and fold homology," *J Mol Biol*, vol. 323, pp. 387–406, Oct 2002.

[40] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database ef-site," *Protein Sci*, vol. 12, pp. 1589–95, Aug 2003.

[41] R. B. Russell, "Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution," *Journal of Molecular Biology*, vol. 279, no. 5, pp. 1211–1227, 1998.

[42] G. J. Kleywegt, "Recognition of spatial motifs in protein structures.," *Journal of Molecular Biology*, vol. 285, pp. 1887–1897, Jan 1999.

[43] C. Ferrer-Costa, H. P. Shanahan, S. Jones, and J. M. Thornton, "Hthquery: a method for detecting dna-binding proteins with a helix-turn-helix structural motif," *Bioinformatics*, vol. 21, pp. 3679–80, Sep 2005.

[44] B. Chen, V. Fofanov, D. Kristensen, M. Kimmel, O. Lichtarge, and L. Kavraki, "Algorithms for structural comparison and statistical analysis of 3D protein motifs.," in *Proceedings of Pacific Symposium on Biocomputing*, vol. 334, p. 45, 2005.

[45] B. Chen, D. Bryant, V. Fofanov, D. Kristensen, A. Cruess, M. Kimmel, O. Lichtarge, and L. Kavraki, "Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction.," *J Bioinform Comput Biol*, vol. 5, no. 2a, pp. 353–82, 2007.

[46] B. Y. Chen, D. H. Bryant, A. E. Cruess, J. H. Bylund, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Composite motifs integrating multiple protein structures increase sensitivity for function prediction," in *Proc. of the Sixth*

Annual Intl. Conf. on Computational Systems Bioinformatics (CSB2007), pp. 343–355, Imperial College Press, 2007.

[47] L. Holm and J. Park, "Dalilite workbench for protein structure comparison," *Bioinformatics*, vol. 16, pp. 566–7, Jun 2000.

[48] O. C. Redfern, A. Harrison, T. Dallman, F. M. G. Pearl, and C. A. Orengo, "Cathedral: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures," *PLoS Comput Biol*, vol. 3, p. e232, Nov 2007.

[49] I. Shindyalov and P. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Engineering Design and Selection*, vol. 11, no. 9, pp. 739–747, 1998.

[50] T. Madej, J. F. Gibrat, and S. H. Bryant, "Threading a database of protein cores," *Proteins*, vol. 23, pp. 356–69, Nov 1995.

[51] P. I. de Bakker, A. Bateman, D. F. Burke, R. N. Miguel, K. Mizuguchi, J. Shi, H. Shirai, and T. L. Blundell, "Homstrad: adding sequence information to structure-based alignments of homologous protein families," *Bioinformatics*, vol. 17, pp. 748–9, Aug 2001.

[52] R. D. Finn, J. Clements, and S. R. Eddy, "Hmmer web server: interactive sequence similarity searching," *Nucleic Acids Res*, vol. 39, pp. W29–37, Jul 2011.

[53] C. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, "PROSITE: A documented database using patterns and profiles as motif descriptors," *Briefings in Bioinformatics*, vol. 3, no. 3, pp. 265–274, 2002.

[54] H. Thompson J D, "GibsonTJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, no. 4, pp. 673–4.

[55] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–7, 2004.

[56] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc Natl Acad Sci U S A*, vol. 89, pp. 10915–9, Nov 1992.

[57] R. Schwartz, M. Dayhoff, and B. Orcult, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352, 1978.

[58] C. Schalon, J.-S. Surgand, E. Kellenberger, and D. Rognan, "A simple and fuzzy method to align and compare druggable ligand-binding sites," *Proteins*, vol. 71, pp. 1755–78, Jun 2008.

[59] J. Platt, "Fastmap, MetricMap, and Landmark MDS are all Nyström algorithms," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pp. 261–268, Omni Press Madison, WI, 2005.

[60] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.

[61] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[62] L. Xie, L. Xie, and P. E. Bourne, "A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to

genome-based drug discovery.," *Bioinformatics*, vol. 25, pp. i305–12, Jun 2009.

[63] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25–9, May 2000.

[64] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, "The PDBbind database: methodologies and updates," *J Med Chem*, vol. 48, pp. 4111–9, Jun 2005.

[65] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist, "The PROSITE database," *Nucleic Acids Res*, vol. 34, pp. D227–30, Jan 2006.

[66] J. W. Torrance, G. J. Bartlett, C. T. Porter, and J. M. Thornton, "Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families," *J Mol Biol*, vol. 347, pp. 565–81, Apr 2005.

[67] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Res*, vol. 35, pp. D198–201, Jan 2007.

[68] B. H. Dessailly, M. F. Lensink, C. A. Orengo, and S. J. Wodak, "LigASite–a database of biologically relevant binding sites in proteins with known apo-structures," *Nucleic Acids Res*, vol. 36, pp. D667–73, Jan 2008.

[69] L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner, and H. A. Carlson, "Binding MOAD (Mother Of All Databases)," *Proteins*, vol. 60, pp. 333–40, Aug 2005.

[70] P. de Matos, N. Adams, J. Hastings, P. Moreno, and C. Steinbeck, "A database for chemical proteomics: ChEBI," *Methods Mol Biol*, vol. 803, pp. 273–96, 2012.

[71] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal, "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information," *Bioinformatics*, vol. 19, no. 1, pp. 163–164, 2003.

[72] M. Bashton, I. Nobeli, and J. M. Thornton, "Procognate: a cognate ligand domain mapping for enzymes," *Nucleic Acids Res*, vol. 36, pp. D618–22, Jan 2008.

[73] I. Halperin, D. S. Glazer, S. Wu, and R. B. Altman, "The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications.," *BMC Genomics*, vol. 9 Suppl 2, p. S2, 2008.

[74] S. Yoon, J. C. Ebert, E.-Y. Chung, G. De Micheli, and R. B. Altman, "Clustering protein environments for function prediction: finding PROSITE motifs in 3D," *BMC Bioinformatics*, vol. 8 Suppl 4, p. S10, 2007.

[75] J. C. Ebert and R. B. Altman, "Robust recognition of zinc binding sites in proteins.," *Protein Sci*, vol. 17, pp. 54–65, Jan 2008.

[76] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[77] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database.," *Nucleic Acids Res*, vol. 36, pp. D281–8, Jan 2008.

[78] K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington, "Homstrad: a database of protein structure alignments for homologous families," *Protein Sci*, vol. 7, pp. 2469–71, Nov 1998.

[79] G. A. Reeves, T. J. Dallman, O. C. Redfern, A. Akpor, and C. A. Orengo, "Structural diversity of domain superfamilies in the cath database," *J Mol Biol*, vol. 360, pp. 725–41, Jul 2006.

[80] J. A. Capra and M. Singh, "Characterization and prediction of residues determining protein functional specificity," *Bioinformatics*, vol. 24, pp. 1473–80, Jul 2008.

[81] F. Pazos, A. Rausell, and A. Valencia, "Phylogeny-independent detection of functional residues," *Bioinformatics*, vol. 22, pp. 1440–8, Jun 2006.

[82] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families.," *Journal of Molecular Biology*, vol. 257, pp. 342–358, Mar 1996.

[83] M. Magrane and Uniprot Consortium, "Uniprot knowledgebase: a hub of integrated protein data," *Database (Oxford)*, vol. 2011, p. bar009, 2011.

[84] R. Peeters, "The maximum edge biclique problem is np-complete," *Discrete Applied Mathematics*, vol. 131, no. 3, pp. 651–654, 2003.

[85] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[86] R. Hamming, "Coding and information theory.," 1980.

[87] S. Madeira and A. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*,

vol. 1, no. 1, pp. 24–45, 2004.

[88] M. A. Fabian, W. H. Biggs, 3rd, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko, A. G. Lai, J.-M. Lélias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, and D. J. Lockhart, "A small molecule-kinase interaction map for clinical kinase inhibitors," *Nat Biotechnol*, vol. 23, pp. 329–36, Mar 2005.

[89] D. Kuhn, N. Weskamp, E. Hüllermeier, and G. Klebe, "Functional classification of protein kinase binding sites using Cavbase," *ChemMedChem*, vol. 2, pp. 1432–47, Oct 2007.

[90] S. L. Kinnings and R. M. Jackson, "Binding site similarity analysis for the functional classification of the protein kinase family," *J Chem Inf Model*, vol. 49, pp. 318–29, Feb 2009.

[91] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer Series in Statistics, 2001.